



3 1761 10374372 0



Digitized by the Internet Archive
in 2023 with funding from
University of Toronto

<https://archive.org/details/31761103743720>

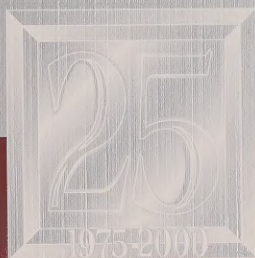
12-001



Government
Publications

266

SURVEY METHODOLOGY



Catalogue No. 12-001-XPB

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2000

•

VOLUME 26

•

NUMBER 1



Statistics
Canada

Statistique
Canada

Canada



SURVEY METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

June 2000 • VOLUME 26 • NUMBER 1

Published by authority of the Minister
responsible for Statistics Canada

© Minister of Industry, 2000

All rights reserved. No part of this publication may be reproduced,
stored in a retrieval system or transmitted in any form or by any
means, electronic, mechanical, photocopying, recording or otherwise
without prior written permission from Licence Services,
Marketing Division, Statistics Canada,
Ottawa, Ontario, Canada K1A 0T6.

July 2000

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa



Statistics
Canada

Statistique
Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is abstracted in The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

MANAGEMENT BOARD

Chairman G.J. Brackstone

Members D. Binder
G.J.C. Hole
F. Mayda (Production Manager)
C. Patrick

R. Platek (Past Chairman)
D. Roy
M.P. Singh

EDITORIAL BOARD

Editor M.P. Singh, *Statistics Canada*

Associate Editors

D.R. Bellhouse, *University of Western Ontario*
P. Biemer, *Research Triangle Institute*
D. Binder, *Statistics Canada*
C. Clark, *U.S. Bureau of the Census*
J.-C. Deville, *INSEE*
J. Eltinge, *Texas A&M University*
W.A. Fuller, *Iowa State University*
M.A. Hidioglou, *Statistics Canada*
D. Holt, *Central Statistical Office, U.K.*
G. Kalton, *Westat, Inc.*
P. Kott, *National Agricultural Statistics Service*
P. Lahiri, *University of Nebraska-Lincoln*
S. Linacre, *Australian Bureau of Statistics*
G. Nathan, *Central Bureau of Statistics, Israel*

D. Norris, *Statistics Canada*
D. Pfeffermann, *Hebrew University*
J.N.K. Rao, *Carleton University*
L.-P. Rivest, *Université Laval*
I. Sande, *Telcordia Technologies*
F.J. Scheuren, *The Urban Institute*
R. Sitter, *Simon Fraser University*
C.J. Skinner, *University of Southampton*
E. Stasny, *Ohio State University*
R. Valliant, *Westat, Inc.*
J. Waksberg, *Westat, Inc.*
K.M. Wolter, *National Opinion Research Center*
A. Zaslavsky, *Harvard University*

Assistant Editors J.-F. Beaumont, P. Dick, H. Mantel, B. Quenneville and D. Stukel, *Statistics Canada*

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their manuscripts in either English or French to the Editor, Dr. M.P. Singh, Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Four nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of Survey Methodology (Catalogue no. 12-001-XPB) is CDN \$47 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN \$12 (\$6 \times 2 issues); Other Countries, CDN \$20 (\$10 \times 2 issues). Subscription order should be sent to Statistics Canada, Dissemination Division, Circulation Management, 120 Parkdale Avenue, Ottawa, Ontario, Canada K1A 0T6 or by dialling 1 800 700-1033, by fax 1 800 889-9734 or by E-mail: order@statcan.ca. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, the Statistical Society of Canada and l'Association des statisticiennes et staticiens du Québec.

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Volume 26, Number 1, June 2000

CONTENTS

In This Issue	1
G. KALTON	
Developments in Survey Research in the Past 25 Years	3
D.R. BELLHOUSE	
Survey Sampling Theory Over the Twentieth Century and its Relation to Computing Technology	11
B.A. BAILAR	
The Past is Prologue	21
C.T. ISAKI, J.H. TSAY and W.A. FULLER	
Estimation of Census Adjustment Factors	31
R. LACHAPELLE and D. KERR	
Census Coverage Error: A Demographic Evaluation	43
M. FEDER, G. NATHAN and D. PFEFFERMANN	
Multilevel Modelling of Complex Survey Longitudinal Data With Time Varying Random Effects	53
L.-P. RIVEST and E. BELMONTE	
A Conditional Mean Squared Error of Small Area Estimators	67
J. SHAO	
Cold Deck and Ratio Imputation	79
S.K. THOMPSON and O. FRANK	
Model-Based Estimation With Link-Tracing Sampling Designs	87
A. THÉBERGE	
Calibration and Restricted Weights	99
W.C. LOSINGER, L.P. GARBER, B.A. WAGNER and G.W. HILL	
A Cautionary Note on Adjusting Weights for Nonresponse	109
J. P. SHAFFER	
Local Unconditional Best Linear Unbiased Estimators: Applications to Survey Sampling	113

In This Issue

This issue of *Survey Methodology* continues the celebration of 25 successful years that was marked by the publication of the December 1999 issue. The first seven papers of this issue, by prominent statisticians working in survey methods, were invited to help mark this occasion but were not included in the December issue due to space limitations. I would like to extend a special word of thanks to all of the authors who helped to make these two celebration issues so special and memorable.

To start off this special issue Kalton reviews developments in survey research over the past 25 years since *Survey Methodology* first started publishing. He first describes developments in survey taking as a profession, specifically the rise of specialist journals and professional associations for survey methodologists, as well as the international and multidisciplinary aspects of the profession. He then reviews developments in survey methods including questionnaire design, data collection, non-sampling errors, sampling methods and estimation. Finally he discusses the rise in importance of panel surveys and international surveys, administrative data sources and analysis of survey data.

Bellhouse traces the parallel developments in survey taking and computing over the twentieth century. He first describes the interaction between census taking and the early development of computing machines and digital computers. Later, developments in scientific computation lead to the use of more sophisticated statistical methods and models. He concludes his story with discussions of the development of statistical software for surveys and of model-related methods.

Bailar discusses the role of statistics in census taking, with particular emphasis on errors in census counts due to census errors of various sorts and adjustment of census counts using sample based estimates of net undercount. The various sources of errors in censuses are described. Use of statistical methods for census evaluation, quality control in census processing, and imputation is also discussed. Using a model for census bias and variance, the potential efficacy of census adjustment procedures is illustrated.

Isaki, Tsay and Fuller consider estimation of census adjustment factors using data from the 1990 post enumeration survey. Their estimators are based on a components of variance model with a fixed linear predictor and a random effect describing the unknown true adjustment factor for each of 336 post-strata. They consider alternatives based on using an estimate of the full variance-covariance matrix of the direct survey errors of the post-stratum adjustment factors versus using only the diagonal elements. Use of the diagonal elements only can reduce the effects of instability in the estimate of the full variance-covariance matrix. In an empirical comparison they find that a compromise between these two extremes works best. They also restrict the model based adjustment factors so that the estimate of total population matches that obtained from the direct survey estimates of these adjustment factors.

Lachapelle and Kerr present an innovative use of a coverage study to examine the demographic estimates of population. Their approach decomposes the results from Statistics Canada's Reverse Record Check (RRC) to provide an additional source of data that can be compared to the more traditional administrative record based estimates of the components of growth. The objective of this comparison is to identify major sources of error in either the administrative record based or the RRC estimates. They also show how the error of closure can be decomposed into two parts: differences between the RRC and the Census estimates of enumerated population and differences between the RRC and administrative record based estimates of growth.

In their paper Feder, Nathan and Pfeiffermann consider repeated sampling from a hierarchical population. At each fixed time point the population can be described by a two level model; first and second level random effects are then allowed to evolve stochastically over time. In particular, the case where second level units remain in the sample for only a few occasions, as for example in many labour force surveys, is considered. A two step estimation procedure is proposed. In the first step the two-level model is fit to each time point independently to obtain estimates of the fixed effects. Time series parameters are estimated in the second step. Sampling weights can be incorporated into both steps to account for possibly informative sampling.

Rivest and Belmonte propose measurement of the mean square error of small area estimators conditionally on the realized smoothing model. They propose a natural estimator for this MSE; however, the estimator can be quite unstable when there is a lot of smoothing. They also propose a correction for bias in the case that the distributions of the direct estimators are skewed. Finally

they investigate the properties of their estimator in an Empirical Bayesian context and illustrate their method using undercoverage data from the 1991 Canadian Census.

Shao addresses an important topic - the evaluation of cold deck imputation methods. Since computer technology continues to make it easier to store and access data from previous and related surveys, imputation methods that make use of this auxiliary data will become increasingly important. As a result, Shao takes the first steps in evaluating how various cold deck imputation methods will perform relative to other imputation methods.

Thompson and Frank discuss model based estimation for link-tracing designs. In link-tracing designs, links are followed from one respondent to another. Network sampling and snowball sampling are just two examples. After a general introduction to the area, they present several link-tracing designs. They then present a graphical model for the linked population. Finally they develop likelihood based inference procedures for such populations using data from link-tracing designs.

Théberge attempts to solve the problem of extreme weights due to the calibration estimator by relaxing somewhat the calibration equation requirements. In fact, the problem is one of minimization similar to that encountered in ridge regression. He also reviews other means of restricting weights. He discusses the asymptotic properties of calibrated weights, and provides necessary and sufficient conditions for the existence of restricted weights satisfying the calibration equation. He also outlines a way of formulating the estimation problem by controlling the significance given to the calibration equation, and describes various means of restricting weights that do not rely on the use of a specific distance. Finally, he suggests an estimator having restricted weights that is useful for small domains, and deals with outliers by developing a method similar to that used to handle extreme weights.

Two short notes conclude this issue. Losinger, Garber, Wagner and Hill present a case study in the care that must be taken when adjusting for non response in different waves of a survey. Finally, Shaffer looks at the estimation of regression coefficients using survey data when the assumption of fixed auxiliary variables is relaxed.

You may recall that the December issue of *Survey Methodology* was made available, on an experimental basis, in an electronic format on the Statistics Canada web site. There was also a web based survey to gauge your reactions and preferences with respect to an electronic version of the journal. Although there was quite a bit of interest in an electronic version, it seems that the time is not yet ripe for publishing electronically on a regular basis. We will certainly be reconsidering this option in the near future, and your responses to the survey will help to improve any future electronic version. In the meantime, we will continue to publish a print version of the journal for the foreseeable future.

M.P. Singh

Developments in Survey Research in the Past 25 Years

GRAHAM KALTON¹

ABSTRACT

In recognition of *Survey Methodology*'s silver anniversary, this paper reviews the major advances in survey research that have taken place in the past 25 years. It provides a general overview of developments in: the survey research profession; survey methodology – questionnaire design, data collection methods, handling missing data, survey sampling, and total survey error; and survey applications – panel surveys, international surveys, and secondary analysis. It also attempts to forecast some future developments in these areas.

KEY WORDS: Survey research profession; Survey methodology; Survey applications; Questionnaire design; International surveys.

1. INTRODUCTION

Survey Methodology is celebrating its silver anniversary this year. In recognition of this milestone, this paper aims to review the major developments in survey research over the past 25 years. I should note, however, that for several reasons I shall be somewhat lax in my dating of events. First, there was, of course, no watershed in survey research in 1975. Rather, many of the major developments over the past quarter century built on the foundations laid by earlier work. Second, it takes time for many advances in methodology to be fully accepted and adopted. Third, I am using as my benchmark a text on survey methodology that Sir Claus Moser and I published in the United Kingdom in 1971 (the second edition of *Survey Methods in Social Investigation*, hereafter referred to as *Survey Methods*), so that my time frame actually extends over 30 years or so.

The paper reviews the developments in survey methodology, including questionnaire design, survey sampling, data collection methods, data processing, and survey analysis. Computers will feature prominently in the discussion since they have had a major impact on many, but not all, methodological developments. The paper also reviews the effects of these methodological developments on the practice of survey research, including the growth in panel surveys, international surveys, and secondary analysis. The main emphasis is on population surveys, but some references are also made to establishment surveys. Also, in reflecting my experience, the paper will no doubt have a slant toward work done in the United States. Before turning to developments in survey methods and practice, I will first describe the great expansion that has taken place in the number of surveys being conducted and the emergence of a clearly identified survey research profession.

2. THE SURVEY RESEARCH PROFESSION

Most of the history of survey research is contained in the twentieth century. The field began to take off in the 1930's,

grew considerably during the Second World War, and has grown at a substantial rate ever since. By 1975, surveys of both households and establishments were well established as the means to meet the needs for statistical data of policymakers and researchers on a wide range of subjects, such as manufacturing and trade, agriculture, employment and unemployment, family expenditure, nutrition, health, education, travel, aging, and crime. In addition, surveys conducted by academic and other researchers in sociology, economics, political science, psychology, education, social work and public health, public opinion and election polls, and market research have flourished. The field has continued to expand at a rapid rate in the past 25 years, particularly as more policymakers have learned to appreciate the value of survey data and as advances in survey methods have enhanced the ability of survey researchers to respond to the demands for statistical data. The continuing demand of policymakers for more and more sophisticated data has prompted advances in survey methodology and has also led to the solidification of a broadly based survey research profession.

The rapid growth in survey research has come about in part because of an expansion in the range of topics that are considered suitable for study using survey methods. Adventurous researchers have constantly and successfully challenged the conventional wisdom of their times about the subject matters that surveys could cover. These challenges have continued during the past 25 years so that there are now very few subjects that are ruled out for study in surveys based on valid probability samples. Some of the new subjects of study are sensitive ones, such as sexual behavior and illicit drug use, for which the application of survey methods has required the development of special data collection techniques. Other new subjects have required the incorporation of additional data collection methods, such as medical examinations for sampled individuals, videotaping of teacher-student interactions in classrooms, and placing environmental monitoring equipment in sampled households. Tackling more difficult subject matters has been a constant stimulus to methodological research.

¹ Graham Kalton, Westat, 1650 Research Boulevard, Rockville, Maryland, USA 20850. E-mail: KaltonG1@westat.com.

Prior to 1975 there were no widely distributed specialist journals in survey methodology. Refereed papers on survey methodology were published in a variety of journals. Statistical journals published, and continue to publish, papers mainly on the more statistical aspects of survey research, particularly survey sampling. Journals like *Public Opinion Quarterly* published, and continue to publish, papers on survey methodology. Market research journals publish papers on survey methodology relevant to market research. Journals in various subject-matter disciplines in the social sciences, public health, *etc.*, sometimes publish papers on survey methods relevant to their disciplines. This situation was not ideal since there was no natural outlet for some good papers on survey research methods and because the literature was widely scattered. The introduction of *Survey Methodology* in 1975 and the *Journal of Official Statistics* in 1985, both now well-established journals, has remedied this situation.

Another notable development has been the establishment of professional associations for survey methodologists. For example, the International Association of Survey Statisticians (IASS) was founded in 1975 as a section of the International Statistical Institute; the Section on Survey Research Methods of the American Statistical Association was established in 1978, after being a subsection of the Social Statistics Section from 1974 to 1977; and the Social Statistics Section of the Royal Statistical Society was formed in 1976, initially as the Social Statistics and Survey Methodology Study Group.

In recent years, several of these associations, sometimes together with other associations (particularly the American Association for Public Opinion Research), have collaborated to run international conferences on specific topics in survey methodology. A special feature of these conferences is that many of them have been structured to cover their chosen topics in a comprehensive manner so that they could generate well-rounded texts. This feature was introduced to address the shortage of literature on survey methodology that resulted from the fact that survey methodologists are practitioners with little time to publish. The result has been the production of edited volumes on such topics as panel surveys, telephone surveys, business surveys, measurement errors in surveys, survey quality, and computer-assisted survey information collection.

Many other conferences on survey methodology have also been held in recent years. Some have been organized by government agencies, such as Statistics Canada, the U.S. Census Bureau and the U.S. Federal Committee on Statistical Methodology (also founded in 1975). Others have been organized by professional associations, such as the IASS and the Association for Survey Computing. The proceedings from these conferences, and those of the Section on Survey Research Methods of the American Statistical Association, contribute greatly to the growth in the literature on survey methodology.

Two other aspects of the development of the survey profession deserve comment. One is its internationalism. The international conferences described above have led to publications with authors from many different countries. Although there are cultural differences between countries that need to be taken into account in data collection, research on survey methodology shares a good deal in common across countries. In addition, international surveys are becoming more prevalent, with the need to standardize procedures across countries (see the discussion below). In general, international cooperation in survey research is progressing well, but there is one area where much more could be done. Like the developed countries, the developing and transition countries need statistical data from surveys. However, they often lack the necessary expertise. The IASS, international agencies like the U.N. Statistical Office, a number of government statistical agencies, and a number of other bodies make valuable contributions to training survey researchers from developing and transition countries, but the level of support currently available for this training falls far short of what is needed.

Another noteworthy aspect of the development of the survey research profession is its multidisciplinary nature. As survey research has become established as a profession, it has developed a number of subdisciplines. Thirty or so years ago, a survey methodologist might expect to cover all aspects of the subject, but that is no longer possible at the highest technical level. The statistical level of the techniques of survey sampling and survey analysis used by survey statisticians has advanced greatly, survey methodologists are increasingly using theories and techniques from sociology, psychology, and anthropology, and computer specialists now need to use much more sophisticated methods for data capture and processing than in the past. This inevitable segmentation of survey methodology as the field progresses puts at risk a unified professional identification, particularly since the subdisciplines are each also associated with their own different fields. Given the importance of interdisciplinary collaboration in survey research, mechanisms to foster that collaboration may be needed in the future (see also section 5).

As with the developing and transition countries, the developed countries face a shortage of well-trained survey statisticians and methodologists. There is the need both to attract more people into the profession and to provide more training opportunities for them. There are a few graduate programs at universities and some faculty who specialize in the field, but the numbers are inadequate given the needs. The multidisciplinary collaboration involved in constructing and conducting a survey implies that the training should have a multidisciplinary component, so that the various specialists can communicate effectively with one another. Moreover, the instructors should include persons with practical survey experience. These specifications make it even more difficult for a graduate program in survey methodology to be mounted in most universities. An

alternative approach is that adopted by the Joint Program in Survey Methodology (JPSM) at the University of Maryland, a program set up with U.S. government funds to address the shortage of trained survey researchers in the federal government. The JPSM is built on a collaboration of two universities (the University of Maryland and the University of Michigan) and a private survey research organization (Westat), with important contributions from experts in survey methodology in the government, other organizations, and other universities to support its various graduate programs. In a related approach, the Department of Social Statistics at the University of Southampton and the U.K. Office for National Statistics have recently jointly developed a master's degree program in official statistics, with significant teaching contributions in both survey methodology and other aspects of official statistics being made by government statisticians. The Department is also collaborating with an independent survey research organization (the National Centre for Social Research) in the Centre for Applied Social Surveys, one activity of which is to run short courses in survey methodology.

3. DEVELOPMENTS IN SURVEY METHODS

The computer revolution that began to have a significant impact on survey analysis in the 1960's has been the dominating force behind the advancement of survey methodology over the past 25 to 30 years. The ability to process and analyze survey data much more readily than in the past has supported the use of more advanced statistical methods. It has also contributed greatly to more sophisticated demands from survey data users, stimulating the development of improved methodology for all aspects of the survey process.

The chapter on processing survey data in *Survey Methods* contains a description of punch cards that were widely used 30 years ago for the analysis of survey data, together with a description of unit record equipment (counter-sorters and tabulators) and computers. At that time computers were well on the way to replacing unit record equipment, but they were not routinely available to survey researchers. The computers of the day were large main-frame machines and punch cards were the usual input medium for survey data. Programs for survey analysis were limited in number and in scope. Today, the situation is, of course, totally different, and the impact of this change on survey research is hard to overstate.

It is against this backdrop of the computing explosion that the advances in other aspects of survey methodology should be assessed. The rest of this section briefly outlines what I view to be the significant advances that have been made in the past quarter century in the areas of questionnaire design, data collection, missing data, survey sampling, and total survey error.

Questionnaire design. The critical role of questionnaire design in achieving high-quality survey data has been well

recognized from the early days. While some first-rate research was being conducted on improving questionnaire design in the 1960's and 1970's, the number of researchers involved in tackling this extremely challenging area was very limited. This situation has improved subsequently in large part due to what has become known as the Cognitive Aspects of Survey Methodology (CASM) movement. The CASM movement aims to attract researchers from the cognitive and social sciences to address the difficult problems of formulating survey questions that produce appropriate responses. The attention generated by this movement has created renewed interest in this field.

The CASM movement has not identified ready-made solutions to the problems of response errors in surveys. It would have been unrealistic to expect that all that was needed was the importation of existing theories from cognitive psychology and other disciplines into questionnaire design. What the movement has achieved is greater efforts to tackle the subject from a theoretical perspective. Also, the CASM movement has contributed greatly to more rigorous pretesting of survey questionnaires. Some of the pretesting techniques that have been developed in the past 25 years occurred independently of the CASM movement, but the sustained attention that pretesting now receives owes a great deal to that movement. A direct effect of the CASM movement has been the creation of the so-called "cognitive laboratories" that are now widely used for pretesting questionnaires, using such techniques as "think alouds" and extensive probing. Focus groups – which have a long history in questionnaire design, particularly in market research – are also much more widely used than in the past. In addition, behavior coding is now used widely in pretesting.

An associated development in the past few years has been a more theoretical approach to the design of forms that are to be completed by survey respondents. This research takes account of theories that indicate how individuals approach documents and how they most naturally work their way through them. This important subject received little attention for many years. The current research holds considerable promise for making survey forms much more user friendly, with the hope that this may improve both the quality of the data collected and response rates.

Data collection. *Survey Methods* contains two main chapters on data collection methods, one on mail questionnaires and one on face-to-face interviewing (there is also a chapter on documents and observation). There are only a few minor references to telephone interviewing, in part because of the low level of telephone penetration in the United Kingdom at that time. However, even in the United States where telephone penetration was much higher, back in 1975 many survey researchers had serious doubts about the collection of data for household surveys by telephone, at least for government surveys with major policy implications. That situation has changed considerably. Today, many U.S. government surveys are conducted by telephone.

One concern about telephone surveys is the noncoverage of households without telephones. With telephone coverage in the U.S. currently around 95 percent, the noncoverage of nontelephone households may be considered acceptable for surveys of the general population. However, a sizable number of surveys focus on subpopulations with lower telephone coverage rates, such as the poor; for such surveys telephone noncoverage is a serious concern. Another concern is nonresponse. Nonresponse rates for telephone surveys are appreciably higher than for comparable face-to-face interview surveys, and the gap appears to be widening. In making a choice between telephone and face-to-face modes of data collection, the large cost savings that accrue from the use of telephone interviewing often override the higher response rates achievable with face-to-face interviewing. Nevertheless, the risk of appreciable bias that is associated with high levels of nonresponse in telephone surveys (frequently as high as 40 percent or more, even with determined follow-up efforts) is a serious and often underrated concern. The likelihood of increasing nonresponse rates to telephone surveys raises questions about the role of telephone data collection in the future.

An important advance in data collection methods in recent years has been the introduction of computer-assisted methods, such as computer-assisted personal interviewing (CAPI) and computer-assisted telephone interviewing (CATI). These methods facilitate more complex skip patterns, prevent interviewers from deviating from the specified question sequence, provide for easy insertion of responses from earlier questions (e.g., if a son's name is recorded as "Peter" in answer to one question, "Peter" can be inserted in the wording of a subsequent question), and enable edit checks to be carried out as the interview progresses and corrections made as necessary. By entering the data directly into a computer file, they also permit more timely processing. The development of general purpose programs for CAPI or CATI data collections, including sampling and scheduling, is a complex operation. Several programs are now available for this purpose. Future developments should see more flexible programs and authoring systems that are simpler to apply.

In the past few years, another form of computer-assisted survey information collection has emerged. This is computer-assisted self-interviewing (CASI), of which there are several variants: video-CASI, in which the respondent reads the questions on the computer screen and enters the answers on the keyboard; audio-CASI, in which the respondent listens to questions on headphones connected to a laptop computer and enters the answers on a keyboard; and telephone audio-CASI in which the audio-CASI interview is conducted by telephone, either with the respondent calling into the computer or with the respondent being transferred to the computer interview once the call has been established by a telephone interviewer. All these versions of CASI avoid the respondent-interviewer interactions that apply with other interviewing methods, and may therefore

be particularly useful for collecting data on sensitive issues. They can also be developed in different languages if necessary. The audio variants avoid the requirement that the respondent is literate. These methods have appeared only recently and their use may be expected to expand appreciably in the future.

Some business surveys are now conducted using audio-CASI methods. An advantage to respondents is that they can call in to a toll-free number at a time convenient to them. They then listen to voice-digitized survey questions and enter responses on the keypad of a touchtone telephone. A variant of this methodology is for the respondents to answer verbally, with the responses interpreted using voice recognition techniques. The use of this methodology may increase in the future as voice recognition methods improve.

Another recent development has been the collection of survey data over the Internet. This methodology is particularly attractive for some types of establishment surveys and for surveys of populations of individuals who have access to the Internet and experience in using it. One approach is to send the questionnaire by email, which may be suitable for individuals who have known email addresses (e.g., the employees of a firm with its own network). Another approach is to post the questionnaire at a web site, with respondents using a password to gain access to it. At this time, the Internet is not appropriate for use in surveys of the general population because of the high proportion of persons without ready access to it, the lack of a sampling frame, and likely low response rates. The temptation to collect a large sample of Internet responses to a survey questionnaire in an uncontrolled fashion should be avoided. This approach would simply replicate the errors made with the infamous 1936 Literary Digest Poll.

Missing data. Missing data occur in surveys through total nonresponse, item nonresponse, and noncoverage. During the past 25 years and even earlier, there has been increasing concern that total nonresponse rates have been rising. This trend is hard to document and indeed analyses of trend data from different surveys have led to different conclusions about the existence of a trend. Yet there is common agreement among survey practitioners that it has become more difficult over time to obtain cooperation. Various reasons have been suggested, such as less novelty in participating in a survey, more working people with less leisure time, fear of crime in face-to-face surveys, and the negative effects of telemarketing in telephone surveys, but there are no definitive explanations. Whatever the reasons, greater efforts now need to be made to achieve a high response rate than was the case in earlier times. These efforts include increased numbers of calls to contact respondents, greater efforts in refusal conversion, and the greater use of incentives. In the past decade, a sizeable number of experimental studies have been conducted in face-to-face and telephone interview surveys to test the effects on response rates of

various monetary and nonmonetary incentives and the level of monetary incentives, thus replicating in an interview setting the kinds of studies that were conducted with mail questionnaires in earlier decades.

Noncoverage is a recognized concern in telephone surveys, but it has received less attention in face-to-face interview surveys, and certainly less attention than the problem of nonresponse. Yet the level of noncoverage in face-to-face interview surveys among certain segments of the population (e.g., young black males in the United States) can be high. Moreover, little is known about those not covered, except that they can be expected to be different in many ways from those covered. It is a source of survey error that would benefit from greater attention in the future. Noncoverage is often especially severe when a survey of a rare population (e.g., teenagers) is conducted with sample members being identified through a large-scale screening survey. Given the increasing interest in surveying rare populations, this type of noncoverage warrants particular attention.

Twenty-five years ago, item nonresponse was generally handled by simply dropping the cases from the analysis in question, for example computing percentage distributions for the subset of cases with acceptable responses. In essence, the implicit assumption being made was that the item nonresponses were missing completely at random (MCAR). Although that practice is still applied in many surveys, increasingly some form of imputation is being used to assign values for the missing responses in a manner that takes account of responses to other survey questions. This process replaces the often untenable MCAR assumption by a missing at random (MAR) assumption, that is that the item nonresponses are missing at random conditional on the auxiliary variables used in the imputation. Although imputation methods were occasionally used 25 years ago, most of the substantial literature on the subject has appeared since 1975. Current methods rely heavily on the computer power that is now available. Imputation remains an area of active research with two main foci: the development of imputation methods that maintain the covariance structure of the survey data set, taking into account that nearly all of the survey variables may be subject to item nonresponse; and the computation of variance estimates for survey estimates that are based on data some of which are imputed (see the discussion below).

Data editing is closely related to imputation. It has also experienced significant advances in recent years, taking advantage of increased computing power to develop more complex editing procedures than could have been employed in the past. Like imputation, editing is the subject of much current research interest and further developments can be expected.

The growth in computing power is also a major factor in the development and widespread use of weighting adjustments for nonresponse and noncoverage. Weighting class adjustments for nonresponse and noncoverage (poststratification) were applied when unit record equipment was used

for survey analysis, but the methods were necessarily relatively simple. Now, more complex weighting class methods and calibration methods incorporating numerous auxiliary variables are widely used, often after exploratory analyses have been conducted to identify appropriate auxiliary variables.

Survey sampling. The main methods of sample design (e.g., stratification, multistage sampling, sampling with unequal probabilities) were developed in the early years and were described in textbooks that appeared in the 1950's. The developments in the past quarter century have been refinements and extensions of these methods, for example to random digit dialing (RDD) sampling for telephone surveys. Here again, the ability of the computer to process large volumes of data in census files and other large sampling frames has enabled survey statisticians to construct more efficient sample designs than in the past.

One area of research in recent years has been on methods for sampling rare populations, either in a special survey or by oversampling in a general survey. This interest is part of the extension of survey demands to provide results for many different domains, including small domains such as racial and ethnic minorities, children in poverty, age/sex groups, and geographical subdivisions (see also the reference to small area estimation below). The aim of the research is to develop efficient sample designs and data collection methods for sampling such domains in situations where special frames for those domains are unavailable. Since the demands for domain results continue to grow, ways to survey rare populations in a cost-effective manner will continue to be sought.

In the 1970's, the design-based mode of inference that is generally adopted with sample surveys was strongly challenged by those who argued that it should be replaced by the model-dependent methods used in the rest of statistics. That debate has waned, and the design-based framework remains in place (see the further discussion below). In this context, the terminology should be clarified: from early on, the design-based mode of inference incorporated the use of models in improving the precision of survey estimates (e.g., regression estimates), but the estimates remained consistent under that mode of inference irrespective of the validity of the model. Thus, the procedures are model-assisted as distinct from model-dependent. The suitability of model-dependent estimates depends on the validity of the model (or the robustness of the estimates to model failure). The computing developments of recent years have facilitated the greater use of models, and of more complex models, within the design-based model-assisted framework of inference.

These remarks should not be interpreted to imply that model-dependent methods have no place in survey research. On the contrary, the methods for handling missing data described above are necessarily model-dependent. Model-dependent methods are also used increasingly in producing estimates for small domains (generally small geographic

areas). Such methods are needed when the sample sizes in the domains are too small (they may often be zero) to produce design-based estimates of adequate precision. In this situation, small area estimates may be produced by borrowing strength from survey data for other areas or time periods through a statistical model that relates the survey data to other, generally administrative, data. The rapid growth in social programs that distribute funds to small geographic entities has led to a substantial demand for up-to-date small area estimates. As a result, small area estimation has become a major area of research activity in recent years, and is likely to remain so in the years to come.

Variance estimation for estimates from complex sample designs has been another major area of development in the past quarter century. Methods based on Taylor's series approximations and replication methods were being used in the 1960's, but they were not routinely applied and were largely confined to research studies. This situation has changed dramatically as a result of the increases in computing power and the development of a number of computer packages for the computation of sampling errors for estimates from complex (typically stratified multistage) sample designs. It is now fairly common practice to compute sampling errors routinely in analyzing survey data.

A notable development in recent years has occurred in the area of the application of analytic models to survey data. This area is one where there remains a debate about the choice between a design-based and model-dependent mode of inference. Within the design-based framework, there have been both theoretical advances in the application of regression models, categorical models, survival models, multilevel models, *etc.*, with survey data and in software for computing variances for these models. At present, survey analysts often conduct their exploratory analyses using the greater flexibility of standard statistical packages, and compute the design-based variances using survey sampling variance estimation software only at the final stages of their analyses. In the future, survey sampling variance estimation procedures should become more fully integrated into standard packages.

An area of much current research activity is the computation of variance estimates for survey estimates that are based on responses some of which are imputed. One approach is the application of multiple imputation procedures to complex sample designs, an application that makes strong use of current computing power. Other methods are being developed under the standard design-based mode of inference (necessarily with model assumptions). The future may see the incorporation of these methods into the survey sampling variance estimation programs so that they can be readily applied.

Total survey error. The preceding discussion has treated the various components of the survey process individually. A well-designed survey, however, is the blending together of the components into an effective package taking cost considerations into account. The last 25 years have seen a

firmer recognition of the issue, with heightened attention to the concepts of total survey error and total survey design. With constrained resources, a survey design reflects trade-offs between, for example, sample size, the extent of non-response conversion undertaken, questionnaire length, and the quality of data obtained by different modes of data collection. In analyzing survey data, the quality of the estimates should properly be assessed in terms of the total survey error from all sources, not just sampling error. For both design and analysis, detailed information is needed on the various sources of error and their effects on the survey estimates. Moreover, since surveys are multipurpose studies, with many different analytic goals, the information requirements are extensive. The rapidly growing literature on survey errors from different sources is helpful for addressing total survey error and total survey design within cost constraints, but more studies are still needed.

The total survey error and total survey design concepts are most readily applied to repeated surveys. Information on error sources can be accumulated from one round to the next and can then be used to determine priorities for where improvements in the survey methods are most needed. One use of the quality profiles that provide integrated accounts of what is known about the error sources in a survey (see the discussion below) is to guide the choice of priorities for methodological improvements.

4. OTHER DEVELOPMENTS

This section reviews a number of areas of survey research in which important developments have occurred in the past 25 years, other than the strictly methodological areas discussed in section 3. The set is not intended to be an exhaustive one. It includes only areas that I consider to have undergone major change.

Panel surveys. The benefits of longitudinal data obtained from panel surveys have long been recognized, and panel surveys were being conducted in the 1940's and 1950's. At that time, however, the complexities of creating longitudinal data sets, combining the data collected in different waves, were severe. Panel surveys were often mostly analyzed only cross-sectionally, and this was a major source of criticism of the method. Today, the advances in computing and also in techniques for longitudinal analysis have changed the situation dramatically. Nevertheless, the complexities of longitudinal data, especially the problem of missing data, remain. Longitudinal methods of analysis are now widely used, although many panel surveys are still analyzed mostly cross-sectionally, with too little attention to the wide range of issues that their longitudinal data could illuminate.

There has been an enormous growth in panel surveys in the past 20 years, covering a wide range of subjects, including education, labor force transitions, health, and voting behavior. Panel surveys of household economics, modeled on the University of Michigan's Panel Study of

Income Dynamics that began in 1968, have become popular and are now being conducted in a sizeable number of countries. There are also panels like Statistics Canada's Survey of Labour and Income Dynamics and the U.S. Census Bureau's Survey of Income and Program Participation that use similar approaches.

It seems likely that the use of panel designs will increase even more in the future. The challenge is to make full use of the longitudinal data produced, since the analytic potential of a panel survey increases exponentially with the number of waves of data it collects. In addition, the significant advances in techniques for longitudinal analysis being made by biostatisticians and others provide the tools for more sophisticated analyses than in the past. Many skilled analysts are needed if the data collected in a panel survey are to be fully analyzed. The growth of secondary analysis (see the discussion below) holds promise for fuller use of panel survey data in the future.

International surveys. The last 25 years have seen the emergence of international surveys of various kinds, ranging from surveys promoted by international agencies to the coordination of independent country surveys to provide cross-national comparisons. A major breakthrough in this area came with the World Fertility Survey (WFS), which conducted surveys in 42 developing countries and 20 developed countries during the period 1974-1982. The WFS not only collected valuable data on fertility, but in many countries it also provided technical assistance in survey research that helped to develop an infrastructure of survey taking. The ongoing Demographic and Health Survey began shortly after the end of the WFS and to date has conducted surveys in more than 50 countries.

Education has been the subject of a number of international surveys including, for example, the Third International Mathematics and Science Study (41 countries in 1995) and its replication (40 countries in 1999); the Programme for International Student Assessment (about 30 countries in 2000); the Second Civics in Education Study (about 20 countries in 1999); the IEA Reading Literacy Study (about 30 countries in 1991). The ongoing International Adult Reading Literacy Survey is collecting comparable information about literacy levels of adults in a number of countries around the world. Two examples of other internationally organized survey designs are the Multiple Indicator Cluster Survey from UNICEF and the Social Dimensions of Adjustment Integrated Survey from the World Bank. A related activity is the coordination of surveys in the European Union by Eurostat. An example of cross-national collaboration on surveys is provided by the International Social Survey Programme, a continuing annual survey program on social science topics that now has 33 member countries.

The development of international survey programs has occurred for two separate reasons. One is the growing interest in the comparison of survey results across countries.

The other is to assist countries, particularly developing and transition countries with limited survey experience, in the conduct of surveys that will provide important data for planning purposes. Considerable expansion in international survey activity can be expected in the future for both of these reasons.

Linkages to administrative data. The increases in computing power and the resultant ability to conduct more sophisticated analyses have led to a demand for more data on the sampled units. Analysts want to answer more complex questions than was the case in the past and some of the data they need may not be readily collectable in a survey, at least with the required level of quality. Even if the data were collectable, the collection could create excessive respondent burden. This situation has led to the search for alternative sources for the data, with data taken from those sources then being linked to the survey responses. Thus, for example, tax records might provide valuable earnings histories for sampled individuals over a timespan for which the respondents could not provide the data, or medical records might provide the amounts of medical expenses paid directly by insurers that are unknown to the respondents. These kinds of linkages have been made much more feasible by the significant expansion in the number of administrative record systems now available in electronic form.

There has been considerable interest in linking administrative record data to social survey data in recent years and a number of surveys have made such linkages. However, there are generally significant problems to overcome in gaining access to administrative data and serious concerns about protecting the survey respondents' privacy. These issues have severely limited the use of administrative record linkages in household surveys to date. Despite the substantial potential benefits of such linkages, it is not clear to what extent these barriers can be overcome.

In contrast, administrative data have become a key element in the conduct of economic surveys and censuses and, in a number of cases, they have replaced the data that used to be collected from respondents. The result has been a substantial decrease in respondent burden, improved data quality, more timely reporting, and reduced costs.

Secondary analysis. The increases in computing power, the increasing numbers of surveys being conducted, and the increased sophistication of the data collected in surveys have all stimulated a major growth in the secondary analysis of survey data. Public-use files are now more routinely made available, sometimes through survey data archives, to enable secondary analysts to conduct their own analyses, thus permitting survey data to be more thoroughly analyzed. Associated with this activity, increased attention has been needed to protect the survey respondents' confidentiality and to ensure that data files released to secondary analysts are not used to breach confidentiality. With secondary analysis undoubtedly continuing to expand in the future,

continued attention will need to be given to ways to release survey data in a manner that protects respondents but does not seriously curtail the range of analyses that can be conducted.

Survey quality. Increasing attention is being given to different aspects of survey quality. In the past few years, a number of survey organizations have become interested in survey process quality, applying the ideas of total quality management to survey processes. Greater attention than in the past is being given to quality taken in the broad sense to include the accuracy of the estimates produced, relevance, timeliness, accessibility and cost-efficiency and in the narrower sense of accuracy alone. Users of survey estimates and secondary analysts of survey data need to be informed about the overall quality of the survey data, including sampling errors, nonresponse and noncoverage, response errors and processing errors. While this need has long been recognized, current practice in reporting survey quality is often seriously deficient. There are signs that more attention is now being given to this area. The introduction of quality profiles that provide full and integrated reports on the quality of the data in ongoing surveys is an important contribution.

5. CONCLUDING REMARKS

This section attempts to predict some major considerations for survey research in the next 10 to 20 years. The computer revolution that has transformed the nature of survey research over the past 25 years is still in progress, and further developments can be expected in many aspects of collecting, processing, and analyzing survey data. The telecommunications industry is also in a state of rapid innovation, and the changes are likely to affect the ways that survey data are collected. It seems likely that greater use will be made in the future of mixed-mode designs, taking advantage of new modes for respondents with access to them (*e.g.*, the Internet) and using conventional modes for other respondents. Thus the effect of mode on survey responses will continue to be an important concern.

In general, it seems probable that the demand for survey data will continue to grow rapidly as more policy analysts learn to take advantage of survey data. Increasingly, survey estimates will be needed for small domains, especially small geographic domains, as policymakers target their programs

to special population subgroups. Currently, most of the demand for survey data comes from central governments; in the future the demand from provincial and local governments may expand. The difficulty here is that surveys cost almost as much for small populations as for large ones. Local governments may therefore often be unable to afford the cost of a survey unless inexpensive methods can be found.

The major concern for the future of survey research is that respondents' willingness to participate in surveys may continue to decline, and that increased efforts in data collection will not fully counteract this effect. Thus, response rates will fall. This comment is of particular salience for telephone surveys, where nonresponse rates are already high. A significant increase in telephone nonresponse rates could even lead to the demise of telephone data collection for household surveys.

Finally, the next decade or so may well see the emergence of a new and different professional society for survey researchers that more broadly represents the interests of all members of the profession. Since survey sampling was at the forefront of the developments of survey research in the early years, survey research has strong ties with statistical societies. However, those ties tend to concentrate on survey statistics. There are also ties with societies for public opinion research, market research, and various subject matter disciplines, such as sociology and psychology, primarily for survey researchers who deal with the nonsampling aspects of survey research. Similarly, there are ties with computing societies for those working on survey computing. As yet, however, there is no society that aims to bring survey researchers of all disciplines together. The years to come may see the creation of such a society to promote exchanges across the different disciplines and thereby help to advance the field. Were such a society to be formed, it would not affect the need for the current ties that survey researchers have with statistical and other societies. Survey researchers need to keep in touch both with the developments taking place in survey research broadly and also with the developments in their own disciplines.

ACKNOWLEDGEMENTS

I am grateful to Joe Waksberg and Dan Levine for helpful suggestions in the preparation of this paper.

Survey Sampling Theory Over the Twentieth Century and its Relation to Computing Technology

D.R. BELLHOUSE¹

ABSTRACT

Computation is an integral part of statistical analysis in general and survey sampling in particular. What kinds of analyses can be carried out will depend upon what kind of computational power is available. The general development of sampling theory is traced in connection with technological developments in computation. What is possible in theory is only practicable with the proper computing technology. At the same time new developments in technology can motivate new areas of theory to investigate. One hundred years ago, it was the requirements of statisticians that spurred on technological development. Although theoretical developments in sampling theory have often run ahead of computational capabilities, it is now the case that survey statisticians are now followers of computing technology that has been motivated by others instead of acting as the catalyst that leads to technological change.

KEY WORDS: Analysis of survey data; Digital computers; Punch cards; Scientific programming; Statistical software; Survey data analysis; Survey estimation.

1. INTRODUCTION

There are several ways to approach the history of survey sampling. Two are very tempting, but will not be followed here. The first is to examine sampling in the context of the history of ideas -- who formulated them and then how and why they are formulated, defended and discarded or supplanted. With respect to the personalities, it is not necessarily the one who espouses the idea first who is given prominence but the one who promotes it the best or the one who can best put the idea into practice. The approach of the history of ideas has been followed to a certain extent by Kruskal and Mosteller (1980) and Bellhouse (1988) who examined the progression of ideas beginning with the espousal of the representative method by Kaier (1897) over censuses combined with the use of randomization in surveys by Bowley (1906). The whole story of the debates over the foundations of sampling falls directly under this approach. From this debate, which was initiated by Godambe (1955), has emerged the continuing question of when to use models in sampling design and estimation. A second way to approach the history of sampling is to look at sampling theory as a branch of mathematics and then to fit this development into the general pattern of how research in mathematics evolves. Complicating this is that there are several approaches to how mathematics evolves, as discussed in Gillies (1992). One approach is to note that periodically there are results which seem to open up new areas of research while other areas become seemingly complete or "fished out" for new research ideas. Emerging areas of research often attract several talented researchers to work on these new problems and away from other potential research problems. This has its parallels in sampling. Hansen and Hurwitz (1943) obtained results on sampling

with probability proportional to size and with replacement. Then Horvitz and Thompson (1952) extended this idea to sampling without replacement. The basic problem in unequal probability sampling without replacement is to find a sampling design that yields the desired inclusion probabilities. This resulted in several papers on the subject culminating in the review monograph by Brewer and Hanif (1983). Lately, very few papers are written to promote new without replacement sampling designs that result in inclusion probabilities proportional to a size variable. However, statistics and survey sampling cannot be equated to pure mathematics. Much of statistical research is motivated by practical problems in data interpretation and analysis not by abstract ideas.

In view of the explosion of technology over the 20th century, I chose another approach. This is to view the history of sampling over the 20th century as the history of the interplay between ideas that have been put into practice and computing technology that has defined the limits of practice or that has encouraged ideas for new developments in practice. The development of sampling methods may be categorized by the intersection of two strands: the use of surveys for descriptive and analytic purposes, and whether or not hypothetical models should be used.

2. BEGINNINGS: THE FIRST HALF OF THE TWENTIETH CENTURY

The first two major breakthroughs for survey sampling, one in the formulation of a statistical concept and the other in the development of technology, occurred at the end of the nineteenth century. Both breakthroughs faced some initial opposition or apathy, the idea more so than the technology,

¹ D.R. Bellhouse, Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario, N6A 5B7, Canada

but both prevailed and were developed further. These breakthroughs were: (1) Kaier's (1895/6, 1897, 1905) espousal of sampling through a "representative method" over attempts at complete enumeration for social surveys, and (2) the development of punch card machines for data processing by Hollerith (1894). Both breakthroughs were directly related to survey or census work. This was the first and last time that survey or census issues inspired major technological innovation. From then on, survey sampling has adapted itself to the available technology.

Kaier's idea was to get a sample that was an approximate miniature of the population. Through sampling, more detailed information could be obtained and more specialized studies could be carried out, all at a fraction of the cost of a census. The idea initially met with opposition and it took upwards of a decade for his ideas to be accepted.

The development of machinery by Herman Hollerith for data processing came directly out of the needs of the U.S. Bureau of the Census and the encouragement of the Bureau's Director of Vital Statistics, John Shaw Billings. The events that led to this development are described by Willcox (1926):

"While the returns of the Tenth Census [1880] were being tabulated at Washington, Billings was walking with a companion through the office in which hundreds of clerks were engaged in laboriously transferring items of information from schedules to the record sheets by the slow and heart-breaking method of hand tallying. As they were watching the clerks he said to his companion, 'There ought to be some mechanical way of doing this job, something on the principle of the Jacquard loom, perhaps, whereby holes on a card regulate the pattern to be woven.' The seed fell on good ground. His companion was a young talented engineer in the office who first convinced himself that the idea was practicable and then that Billings had no desire to claim or use it. Thereafter he devoted the bulk of his life with great ultimate profit for himself and the world to ripening the invention and securing its adoption. I have no need to describe or eulogize Hollerith machines."

A full description of the development and use of these machines for surveys is given in Mandeville (1946). Hollerith's machine was applied to processing the 1890 U.S. census. While the 1880 census took over seven years to complete, the 1890 census was finished by early 1895. The Bureau used 180 tons of cards that were processed at a speed of 6,900 cards per 6½-hour day. Not only did the machine save time, it also significantly reduced tabulation errors. The punched card machine was used to process the 1891 Census of Canada, but it did not see early use in the censuses of the United Kingdom and the rest of the British Empire. It was felt that the level of detail required in these

censuses did not justify the use of a Hollerith machine since the time saved by the machine would be balanced by the time taken to punch the card (Hooker 1894). In a paper on census taking Baines (1900) expressed a preference for manual over machine tabulation, especially when labour was cheap. Despite these initial misgivings, improvements to the machine continued and the use of the Hollerith machine for statistics became highly developed by mid-century. Hartley (1946) demonstrated the most sophisticated use of these punched card machines for statistical analysis. This included the calculation of moving averages and serial correlations as well as the solution of simultaneous equations on Hollerith machines.

After these near simultaneous and unrelated innovations in ideas and technology, theory ran ahead of practice for the next 50 or 60 years. Theoretical developments in sampling continued through the first half of the century. Out of discussions over the path to follow in the "representative method," Bowley (1926) put together a monograph describing all the known theoretical results in sampling under random selection and under purposive selection. In addition, he developed the theory for stratified sampling under proportional allocation. The triumph of randomization over purposive selection was due to Neyman (1934) who showed why randomization gave a more reasonable solution to sampling problems than purposive selection. Although not the first to do so, he also developed optimal allocation strategies for stratified sampling. Prior to the middle of the century the last major development, in terms of sampling design with accompanying estimates and variance estimates, was the concept of unequal probability sampling introduced by Hansen and Hurwitz (1943).

The practical implementation of these theoretical results was limited to relatively small-scale surveys. The analyses for most surveys used calculators, either electric ones such as those manufactured by Friden, Marchant or Monroe, or hand calculators operated by turning a crank such as the Brunsviga used by Pearson and the Millionaire used by Fisher. Since the labour in the analysis increased significantly with the sample size, standard errors were seldom calculated, and when calculated the correct formulas were seldom applied. Bowley (1936) describes a typical situation showing the infrequency of standard error calculations:

"Tabulation is usually a dull and tedious job, but there is a certain interest in watching the entries accumulating in a cross table and seeing the gradual growth of continuity out of randomness. When the results take the form of a frequency curve, and especially if we have reason to expect a normal curve and find it, we have good reason to suppose that we have measured satisfactorily a real entity. Thus the distribution of price changes or their logarithms on a normal scale gives a great deal of support to the validity of an index number. In such cases the computation of standard error is reasonable."

Box and Thomas (1944) describe a survey of approximately 4,500 respondents stratified by the industry in which they worked. The standard errors, when presented, were calculated using the formula for simple random sampling. A decade later Deming (1956) noted:

“Although the possibility of showing a valid standard error is by definition a feature of any probability sample, it is a fact that results of probability samples have too often appeared in the past without standard errors because of the sheer labor of computation.”

It is within this context that Mahalanobis (1946) suggested the technique of interpenetrating subsamples. This technique, which Mahalanobis developed at the Indian Statistical Institute in the 1930's (Murthy 1967 and Deming 1956), is very simple: two or more independent subsamples are chosen according to the same sampling design. Then the variation between the subsample estimates of the population total provides an unbiased estimate of the variance of the final estimator of the total. Computationally, the method has distinct advantages in the punch card environment where sums are easier to obtain than variances. With interpenetrating subsamples the main computational effort is in finding the subsample estimates that are based on sums only. The Indian Statistical Institute obtained its first Hollerith machine in 1944. Prior to that time, tabulations and other calculations were done by hand. The Institute's Annual Report for 1945-46 published in *Sankhyā* shows the initial unease that always greets technological change and the eventual positive benefits to change. With respect to the introduction of these machines, the report states:

“Contrary to apprehensions among certain sections of workers that the Hollerith machine would to a large extent eliminate manual computations, it was found that new and detailed studies which could not be formerly undertaken could now be handled without difficulty so that the demand for trained computers in the later stages was on the increase. In addition to routine projects undertaken from time to time, special studies such as mechanical solution of determinants, construction of tables, fitting of orthogonal polynomials, etc. were conducted.”

In the United States, Deming (1956), for example, picked up on the general idea and put forward methods of replicated sampling. The U.S. Bureau of the Census used this method for variance estimation. At the Bureau this idea evolved into pseudo-replication, or eventually balanced repeated replication, for variance estimation (McCarthy 1969).

3. THE ADVENT OF THE DIGITAL COMPUTER

The initial development of the digital computer was for military purposes during the Second World War (Ceruzzi 1998). For some years after the war the military continued to play a central role in the advancement of computing. By the 1950's commercial uses were developed for the computer, and this is where sampling practice begins to catch up with sampling theory. The first generation of commercial computers included the UNIVAC followed by the IBM 700 series. These computers contained thousands of vacuum tubes as internal memory. The tubes for the IBM machine were about three inches in diameter and held 1,024 bits of information. The UNIVAC ran at 2.25 MHz and could carry out 465 multiplications per second. For both machines, data were input via punched cards and stored data was on magnetic tape rather than continued use of the punched cards. The 1961 census in the United Kingdom underscores the continuing central role of the military in computing at this point in time. The census was processed on an IBM 705 computer (Benjamin 1961). The computer belonged to the War Office and was used by the Royal Army Pay Corps. The census workers were able to use the computer when not in use by the army. Information was input via cards punched in one location and then taken to the computer in another location.

Although it was not at the forefront of the development of the computer as it had been with the Hollerith equipment, the U.S. Bureau of the Census was central in the initial commercial development of the digital computer. Not only did the Bureau receive the first UNIVAC that was produced, but also some of its employees participated in design decisions for its construction (Ceruzzi 1998 and Hansen 1987). The computer was delivered in March of 1951 and was used for processing the 1950 census. It ran 24 hours a day all week until the task was completed. Once the census work was completed, the computer was used for other censuses and surveys including the Current Population Survey. Technology was now catching up to theory; the computer was now used for better calculation of variance estimates. It also opened up new possibilities, in particular imputation of missing values. With respect to variance estimates Hansen, Hurwitz, Nisselson and Sternberg (1955) comment:

“Until the acquisition of a high-speed electronic computer, the UNIVAC, extensive approximations were introduced into the estimates of variances to avoid computations that would be exceedingly time consuming with the available equipment. The availability of the UNIVAC makes it possible to avoid

most of these approximations. Even with the electronic computer, however, the work of making variance computations would be extremely heavy if variances were computed for all items directly. Approximate methods will continue to be used in the future, but they will be evaluated by more exact computations than have been feasible in the past."

Other statistical organizations followed but at a slower pace. The slow pace in Canada was perhaps due in part to the American experience. A 1956 report to the Dominion Statistician at the Dominion Bureau of Statistics in Canada on the subject of computing at the Bureau of the Census (reported and quoted by Worton 1998) states:

"Subject-matter people ... are not entirely convinced that the UNIVAC system has given them the results which might be expected from a computer system. Undoubtedly UNIVAC has given a great deal of trouble – much of it probably not the fault of UNIVAC at all. Factors such as poor programming, inadequate analysis of the job, inexperienced operating staff, maintenance problems, and even friction between the three operating groups, *i.e.*, the subject matter staffs, the Central Operations Group, and the Central Electronics Unit are reflected in the performance of the UNIVAC system."

The Dominion Bureau of Statistics, now Statistics Canada, obtained its first computer in 1960, an IBM 705. The computer was used to process the 1961 census. As noted already, the British used an army-owned computer to process their 1961 census. In the late 1940's, Mahalanobis was on a list showing interest in obtaining one of the first UNIVACs (Ceruzzi 1998). However, the annual reports of the Indian Statistical Institute published in *Sankhyā* show that the Institute did not obtain a computer until 1956 at which time it received an HEC-2M.

Variance estimation for survey estimates of means, totals and proportions was now feasible for large-scale surveys. Widespread use of this technology now depended on two things – access to a computer, which was an expensive item to buy, and appropriate software to carry out the calculations.

4. SCIENTIFIC PROGRAMMING

Certain kinds of research, and the application of these research results, are possible only with computing. These possibilities expand not only with the expansion in computing power, but also with easier access to the computer's power through programming languages or packaged programs. For several years the most popular scientific programming language was FORTRAN (FORmula TRANslation). This was introduced in 1957 by IBM for its

704 computer. Part of what popularized FORTRAN was the development of the WATFOR (WATERloo FORtran) compiler at the University of Waterloo in 1965. This popular compiler, which was used for teaching purposes, combined with the dominance of IBM in the marketplace made FORTRAN accessible to many students and subsequently to researchers (Ceruzzi 1998). In reporting on the development of his own computer programs for survey research, Yates (1973) shows how pervasive FORTRAN had become over the 1960's. Yates's programs for the computer at Rothamsted Experimental Station were originally written in the late 1950's with code specific to the computer they had. In the mid-1960's the code was written in Extended Mercury Autocode. By the end of the 1960's this code had to be translated into FORTRAN using a machine translator; otherwise it was not usable at any other computer location. The earliest use of FORTRAN in sampling that I can find is in Fan, Muller and Rezucha (1962). These three individuals, all of who worked at IBM, developed algorithms and accompanying FORTRAN code to select simple random samples by computer.

There were two different paths that were followed in the application of FORTRAN programming to survey sampling. One was among statistical agencies or survey research centres and the other was among individual academic researchers. The kind of work followed along each path is strongly correlated with the evolving power of the computer and the dominance of IBM (and hence FORTRAN) in the market. By the end of the 1960's, many institutions had new and more powerful mainframe computers, often one of the IBM 360 series that was originally announced in 1964. Moreover, the software (FORTRAN in particular) remained compatible with machine changes and upgrades, especially for machines in the IBM 360 series (Ceruzzi 1998). The Dominion Bureau of Statistics obtained its first IBM 360 in 1969, while for example the Universities of Manitoba, Toronto and Waterloo obtained their first machines in the years 1966-67 (Day 1971). At the agencies and research centres, various formulae and procedures necessary to survey design and analysis were computerized. For example, Fellegi, Gray and Platek (1967) report that when the Canadian Labour Force Survey was redesigned over 1964-65, sample selection by Fellegi's (1963) method of unequal probability sampling was coded into a FORTRAN routine. From the University of Michigan Survey Research Center, Kish and Frankel (1970) report that they had FORTRAN code for obtaining variance estimates for a variety of statistics including regression coefficients using balanced repeated replication. By the mid-1960's academic researchers began to use the computer via FORTRAN programming to study, numerically or empirically, the sampling theory that they or others had derived. One of the first was Sedransk (1965) who carried out some efficiency comparisons in FORTRAN on an IBM 7074 (marketed by IBM in 1964) for a double sampling scheme. In particular, efficiency comparisons were made between optimal values

for the first and second phase sample sizes and an approximation to the optimal values. The computations involved taking expected values over a trinomial distribution in which several conditions had been imposed. The use of the computer here was to obtain a numerical comparison between exact methods and approximate ones. By the end of the decade a new kind of computer-based research process emerged. Rao and Bayless (1969) and Bayless and Rao (1970) compared several unequal probability sampling schemes by generating all possible samples and calculating the exact finite population mean square error for several real and constructed populations. It then became the norm to carry out extensive empirical studies on any newly proposed estimator or design.

The past 30 years have seen remarkable changes in computing technology. Modern computers are much faster, physically smaller and have much greater storage capacity. The steady increase in computing power and the availability of standard programming languages has allowed survey researchers to expand as well into survey data analysis. This technological change is reflected in developments in sampling theory for variance estimation. From the 1960's to the 1980's there were three basic computerized approaches to variance estimation of complex survey statistics: Taylor linearization (see Woodruff 1971, for early references to its usage), jackknife (first proposed in sampling by Durbin 1959) and balanced repeated replication (McCarthy 1969). The rise of computing power saw a new technique, Efron's (1982) bootstrap, for variance estimation. This new statistical technique, which was contemporaneous with the development of networked RISC (Reduced Instruction Set Computing) workstations running under a UNIX operating system, is highly computer intensive. Over the 1980's RISC workstations gradually replaced most mainframes in research organizations. Near the end of this transition away from mainframes, Rao and Wu (1987) extended bootstrap methodology to variance estimation for smooth statistics under stratified multistage designs.

The most recent software to have an effect on statistical research is the development of computer algebra packages. Although computer algebra has been in existence for some time, it is only in the last decade that it has progressed to the point that it is accessible to many researchers. With computer algebra many complex manipulations can be done automatically and much quicker than by hand and without risk of error. Similar to several other areas of statistics, many of the algebraic manipulations in sampling theory are related to algorithms that generate partitions. Based on the computer algorithms developed by Andrews and Stafford (1993) and Stafford and Andrews (1993), Stafford and Bellhouse (1997) have extended computer algebra techniques to survey sampling theory. Using their methodology, most of the results of so-called classical sampling theory, either existing in the literature or yet to be obtained, can be derived automatically.

5. ANALYSIS OF SURVEY DATA

While steady and substantial progress had been made in research on problems of survey estimation or enumerative surveys over the 20th century, by 1970 little had been accomplished on the analytical aspects of surveys. The terms "enumerative" and "analytical" surveys were coined by Deming in 1950 (Deming 1953). In the same article he also gives a succinct definition:

"Briefly, the enumerative question is how many? The analytic question is why? is there a difference between two classes, and if so, how big are the differences?"

There is an implication in this quotation that the purpose of analytical surveys was for comparisons of domain means. Certainly, throughout the 1960's the understanding of what constituted an analytical survey was often limited to this. Cochran (1963) states:

"In an analytical survey, comparisons are made between different subgroups of a population, in order to discover whether differences exist among them that may enable us to form or to verify hypotheses about the forces at work in the population."

Yates (1960) also focused mainly on domain comparisons in his discussion of analytic surveys. He did, however, discuss regression analysis and the problem of attenuation, but not the problem of general survey weights. Skinner, Holt and Smith (1989) attribute the pioneering work in analytical surveys to social scientists, Paul Lazarsfeld in particular. I will use the theoretical development of regression analysis in complex surveys to illustrate these connections to social science, in this case economics.

One of the earliest studies to take into account the survey weights in regression analysis was by Klein and Morgan (1951). At the time both were at the University of Michigan; Morgan was in the Survey Research Center. At the outset of their paper they state:

"The sample design, the methods of collecting the data, and underlying economic behavior will all contribute to the formulation of the model. The study of data collected in consumer surveys has convinced us that one cannot proceed simply by the application of conventional statistical methods in the estimation of economic relationships because of the existence of some basic difficulties which we classify as follows: (1) weighting of observations, (2) heteroscedasticity, (3) nonlinearities, (4) the choice of alternative economic concepts, (5) errors of observation."

They addressed the first four "basic difficulties" but not the fifth. In their analysis of the approximately 2,300 responses

to the Survey of Consumer Finances, which was a multi-stage sample, Klein and Morgan used the survey weights through weighted least squares estimation of the regression parameters but ignored the clustering effect when it came to variance estimation. They noted that in many cases the use of the survey weights had little effect on the estimates of the regression coefficient estimates but noted that there was a reduction in the estimated variance for the model error. Though Klein went elsewhere, Morgan remained at the Michigan Survey Research Center. Twenty years later, he and another (Lansing and Morgan 1971) gave an overview of the state of the art for the analysis of economic survey data. Not much had changed in terms of the incorporation of the survey design into the analysis. The same is true for other areas of social research; in many cases not even the survey weights were used. In the economics literature debate continued for at least twenty years over whether to use the survey weights in regression analysis; Porter (1973) has several references to this debate.

It was out of this milieu that Kish, who also worked at the Michigan Survey Research Center, initially put forward the concept of the design effect (Kish 1957), which is the measure of increase or decrease in variance over simple random sampling experienced in a survey with a design other than simple random sampling. Design effects have become central to many aspects of the analysis of complex survey data. With respect to regression analysis, Kish and Frankel (1970) studied the design effects in the estimation of regression coefficients. They used balanced repeated replication to obtain their variance estimates. It is not entirely clear in their presentation exactly what regression coefficients they were estimating. Later, the parameters were explicitly spelled out in Kish and Frankel (1974). Specifically, the finite population parameters are what would be obtained in least squares estimation of superpopulation regression parameters when the entire finite population is available. Estimation of these parameters has become one of the standard approaches to regression analysis from complex surveys. Fuller (1975), using Taylor approximations to the variances, put the whole inference process on a solid theoretical foundation by providing limit theorems for the estimates. In addition, he addressed the one problem that Klein and Morgan (1951) ignored: errors in the variables or measurement errors in the independent variables.

Konijn (1962) took a different approach to regression analysis. Under a cluster sampling design, he assumed different simple linear regression models within each cluster. The parameters of interest were weighted averages of regression parameters with the weights given by the cluster sizes. This approach is model-based in the sense that it is the model parameters that are of interest, not a finite population parameter. Konijn's approach was not followed for several years. However, there is now a substantial literature that has grown out of this model-based approach; Pfeiffermann (1993) contains several references.

With regard to the social science origins of survey analysis, there were similar experiences in categorical data analysis. The sociological literature from the 1960's and on contains many examples of categorical data analysis ignoring the sampling design. After Rao and Scott (1981, 1984) developed contingency table and goodness of fit analyses for complex surveys, Rao and Thomas (1988) tried to promote this methodology among sociologists using a review article. A search through citation indexes shows that, although this work has had great impact in the statistical and medical literature, it has had little impact in the sociological literature. The reason for this may be due, in part, to lack of computer software. The most popular software among sociologists, which is SPSS, does not at the moment contain any routines for the analysis of complex survey data. This points to a wider problem: regression, categorical data analysis and other techniques that have been proposed for complex surveys are not widely practicable without the appropriate computer software. Fuller himself tried to respond to this need by developing a packaged program for survey data analysis (Hidioglou, Fuller and Hickman 1980).

6. STATISTICAL SOFTWARE FOR SURVEY RESEARCH

Frank Yates at Rothamsted Experimental Station was the first statistician to develop software for survey research. His work began in the late 1950's (Yates and Simpson 1960). Originally, programs were written that were specific to each survey. This evolved into a general-purpose program by the early 1960's (Simpson 1961). Although it was the first in the field and was available for many years, it never achieved widespread popularity. There are at least four reasons for its general lack of success, reasons that point to the success of other software developers.

- (1) The package was not user friendly. In his obituary of Yates, Dyke (1995) made allusion to this fact. He says:

"Yates believed that the analyst should understand the relevant theory, and so be ready to specify in exact detail what he wanted. Perhaps for this reason the program was not excessively easy to use! But its power and flexibility, and uncluttered clarity of its output were, and are, outstanding."

- (2) It was too expensive for what it did and could not compete with cheaper competitors. Wolter (1985) lists a number of packages that were available in the mid-1980's. At the time the package was twice as expensive as SUDAAN but could do only tabulations, whereas SUDAAN had the additional capability of regression analysis and ratio estimation.

- (3) Marketing is an important factor in the success of a product. Yates appeared to be more interested in tinkering with his product to improve it rather than investing time in marketing it.
- (4) Other than a manual, by 1985 there was no technical support for the package.

Yates was not alone in having software that did not catch on. I had the same experience when I developed variance estimation software based on tree traversal algorithms (Bellhouse 1985). Other than the expense factor (mine was free), my package was a living example for the other three reasons why some software does not fly.

By the early 1970's there were over 40 packaged programs and routines, written mainly in FORTRAN, that would do statistical analyses (Schucany, Minton and Shannon 1972). Of these original packages only two have remained popular in the marketplace, SAS first released in 1970 and SPSS released in the late 1960's.

The survey software that has maintained predominance in the market for several years is SUDAAN developed by B.V. Shah of the Research Triangle Institute (Shah 1978 and 1984). It is marketed well and fully supported by its developer. It was originally accessed as a SAS procedure and has now become a stand-alone package. The tie with SAS was probably one of the reasons for its initial success. Those who were familiar with SAS could easily familiarize themselves with this new procedure, or equivalently the package, so that in a sense it was user friendly. Further, the package has continued to keep pace with survey research. The original program contained routines to calculate standard errors for survey estimates including means, totals, proportions and ratios. This was expanded to include regression analysis in the late 1970's when research on regression in complex surveys was under way. The program now contains routines for regression analysis, logistic regression, categorical data analysis and survival analysis. It has also kept pace with developments in computing machinery. Originally developed on a mainframe computer, the package is now available for use on a PC. It still maintains its links to SAS, although SAS currently has its own survey analysis procedures under development.

Currently, there are several other programs for survey analysis. The most popular among these programs, in addition to SUDAAN, are STATA and WesVarPC. While SUDAAN has been linked to SAS, the future development of WesVarPC, which was originally developed by the research corporation Westat, has been turned over to SPSS. Further, the survey routines in STATA are part of a larger statistical analysis package. As with mergers in the general business world, along with product and service integration, the future trend for survey data analysis packages is to become part of an omnibus statistical package. The development and maintenance of statistical packages, for survey research or for a wider context, is a time-consuming

enterprise requiring a substantial capital investment. This can only be done by a well-financed organization.

SUDAAN, STATA and WesVarPC, along with the software packages GES from Statistics Canada and another named CLAN, have been recently reviewed and evaluated in Bergdahl, Black, Bowater, Chambers, Davies, Draper, Elvers, Full, Holmes, Lundqvist, Lundström, Nordberg, Perry, Pont, Prestwood, Richardson, Skinner, Smith, Underwood and Williams (1999). SUDAAN and STATA have also been evaluated by Cohen (1997). Among three of the packages reviewed (STATA, SUDAAN and WesVarPC), SUDAAN appears to have the most options. For example, Bergdahl *et al.* (1999) note that SUDAAN carries out variance estimation for complex statistics using any one of Taylor linearization, jackknife and balanced repeated replication. WesVarPC covers jackknife and balanced repeated replication, while STATA relies solely on Taylor linearization. So far, none of the packages does variance estimation using the bootstrap. It may just be a matter of time before this technology is incorporated into these packages. For some of its public use sample files, Statistics Canada provides bootstrap variance estimation procedures in SAS code. These procedures, however, are specific to the surveys in question.

7. MODELS IN SAMPLING

Models have come in and out of favour among sampling practitioners. Due to Neyman's (1934) pioneering work, the paradigm of randomization and the randomization distribution was paramount until the 1960's. However, the use of models did not disappear during the intervening years. Cochran (1946), for example, used models to study certain sampling designs and was able to conclude that systematic sampling was a good design to use under certain population structures. The 1960's debate over models arose out of the questioning of the foundations of sampling initiated by Godambe (1955). Since then the use of models has not only crept back in to sampling theory but has flourished substantially.

Since the 1960's the use of models in sampling has gone in several directions. At the same time, the practical and general use of models in survey estimation and analysis is only feasible with high speed computing and the appropriate software. In keeping with the theme I have been following here, I will take a very narrow approach to models by tying their usage to computing technology.

Several model-related methodologies have been computerized, either through the provision of numerical examples to illustrate the use of the methodology or through simulation studies to examine how the methodology works. At the present time there is only one model-related approach that has matured to the point where a general package program is available. This is the model-assisted approach that C.-E. Särndal has taken over several years resulting in

generalized regression estimation or GREG. The bulk of the work is summarized in Särndal, Swensson and Wretman (1992). The work was initially motivated by the debates over the foundations of sampling. Under a model, a best, in some sense, estimator of a finite population parameter can be derived. Those on the side promoting randomization inference pointed out that when the model fails the associated estimator can perform very poorly. The solution propounded by Särndal was to obtain the estimate under the model and then to adapt it in such a way that it would remain consistent and perform adequately under the randomization distribution. It is an attempt to obtain the best of both worlds. Generalized regression estimation, as well as several other estimators, have been programmed into GES, a generalized estimation system developed at Statistics Canada. This SAS-based software is aimed at the descriptive side of surveys rather than the analytic and is described in Estevao, Hidiogrou and Särndal (1995). It is a package that could easily catch on under the right conditions.

8. CONCLUSIONS

Developments in sampling research are inextricably tied to computing and computational methods. Where research is headed will be guided, in part, by computer developments. What the immediate future holds for computing is greater speed and greater storage capacity so that packages can become bigger and more comprehensive. Generally acceptable practices in survey estimation and the analysis of survey data will be determined by the contents of generally available computer packages for survey sampling. On the research methodology side, new methodology will continue to be increasingly computer intensive. One other foreseeable development is the explosion of the internet. As a result of this explosion, several complete survey datasets are now easily available via the web. The extensive testing of new methodology on a variety of real surveys prior to publication of the methodology may soon become the norm.

ACKNOWLEDGEMENT

This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

- ANDREWS, D.F., and STAFFORD, J.E. (1993). Tools for symbolic computation of asymptotic expansions. *Journal of the Royal Statistical Society, B*, 55, 613-628.
- BAINES, J.A. (1900). On census-taking and its limitations. *Journal of the Royal Statistical Society*, 63, 41-71.
- BAYLESS, D.L., and RAO, J.N.K. (1970). An empirical study of stabilities of estimators and variance estimators in unequal probability sampling ($n = 3$ or 4). *Journal of the American Statistical Association*, 65, 1645-1667.
- BELLHOUSE, D.R. (1985). Computing methods for variance estimation in complex surveys. *Journal of Official Statistics*, 1, 323-329.
- BELLHOUSE, D.R. (1988). A brief history of random sampling. *Handbook of Statistics*. (Eds. C.R. Rao and K.R. Krishnaiah) 6, 1-14. Amsterdam: North-Holland.
- BENJAMIN, B. (1961). The 1961 census of population. *Incorporated Statistician*, 11, 130-143.
- BERGDAHL, M., BLACK, O., BOWATER, R. CHAMBERS, R., DAVIES, P., DRAPER, D., ELVERS, E., FULL, S., HOLMES, D., LUNDQVIST, P., LUNDSTRÖM, S., NORDBERG, L., PERRY, J., PONT, M., PRESTWOOD, M., RICHARDSON, I., SKINNER, C., SMITH, P., UNDERWOOD, C., and WILLIAMS, M. (1999). *Model Quality Report in Business Statistics Volume II: Comparison of Variance Estimation Software and Methods*. London: Office of National Statistics.
- BOWLEY, A.L. (1906). Address to the Economic and Statistics Section of the British Association for the Advancement of Science, York. *Journal of the Royal Statistical Society*, 69, 540-558.
- BOWLEY, A.L. (1926). Measurement of the precision attained in sampling. *Bulletin of the International Statistical Institute* 22 (1), 1-62.
- BOWLEY, A.L. (1936). The application of sampling to economic and sociological problems. *Journal of the American Statistical Association*, 31, 464-480.
- BOX, K., and THOMAS, G. (1944). The Wartime Social Survey. *Journal of the Royal Statistical Society*, 107, 151-189.
- BREWER, K.R.W., and HANIF, M. (1983). *Sampling with Unequal Probabilities*, (Lecture Notes in Statistics, Volume 15). New York: Springer-Verlag.
- CERUZZI, P.E. (1998). *A History of Modern Computing*. Cambridge, Massachusetts: MIT Press.
- COCHRAN, W.G. (1946). Relative accuracy of systematic and stratified random samples for a certain class of populations. *Annals of Mathematical Statistics*, 17, 164-177.
- COCHRAN, W.G. (1963). *Sampling Techniques*, (2nd Edition). New York: Wiley.
- COHEN, S.B. (1997). An evaluation of alternative PC-based software packages developed for the analysis of complex survey data. *American Statistician*, 51, 285-292.
- DAY, N. (1971). *Canadian Computer Census 1971*. Toronto: Canadian Information Processing Society.
- DEMING, W.E. (1953). On the distinction between enumerative and analytic surveys. *Journal of the American Statistical Association*, 48, 244-255.
- DEMING, W.E. (1956). On simplifications of sampling designs through replication with equal probabilities and without stages. *Journal of the American Statistical Association*, 51, 24-53.
- DURBIN, J. (1959). A note on the application of Quenouille's method of bias reduction to the estimation of ratios. *Biometrika*, 46, 477-480.
- DYKE, G. (1995). Obituary: Frank Yates. *Journal of the Royal Statistical Society, A*, 158, 333-338.

- EFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia: Society for Industrial and Applied Mathematics.
- ESTEVAO, V., HIDIROGLOU, M.A., and SÄRNDAL, C.-E. (1995). Methodological principles for a generalized estimation system at Statistics Canada. *Journal of Official Statistics*, 11, 181-204.
- FAN, C.T., MULLER, M.E., and REZUCHA, I. (1962). Development of sampling plans by using sequential (item by item) selection techniques and digital computers. *Journal of the American Statistical Association*, 57, 387-402.
- FELLEGI, I.P. (1963). Sampling with varying probabilities and without replacement: rotating and non-rotating samples. *Journal of the American Statistical Association*, 58, 183-201.
- FELLEGI, I.P., GRAY, G.B., and PLATEK, R. (1967). The new design of the Canadian Labour Force Survey. *Journal of the American Statistical Association*, 62, 421-453.
- FULLER, W.A. (1975). Regression analysis for sample survey. *Sankhyā, C*, 37, 117-132.
- GILLIES, D. (1992). *Revolutions in Mathematics*. Oxford: Clarendon Press.
- GODAMBE, V.P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society, B*, 17, 269-278.
- HANSEN, M.H. (1987). Some history and reminiscences on survey sampling. *Statistical Science*, 2, 180-190.
- HANSEN, M.H., and HURWITZ, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, 333-362.
- HANSEN, M.H., HURWITZ, W.N., NISSELSO, H., and STERNBERG, J. (1955). The redesign of the Current Population Survey. *Journal of the American Statistical Association*, 50, 701-719.
- HARTLEY, H.O. (1946). The application of some commercial calculating machines to certain statistical calculations. Supplement to *Journal of the Royal Statistical Society*, 8, 154-183.
- HIDIROGLOU, M.A., FULLER, W.A., and HICKMAN, R.D. (1980). *SUPER CARP*. Ames: Iowa State U.P.
- HOLLERITH, H. (1894). The electrical tabulating machine. *Journal of the Royal Statistical Society*, 57, 678-689.
- HOOKE, R.H. (1894). Modes of census-taking in the British Dominions. *Journal of the Royal Statistical Society*, 57, 289-368.
- HORVITZ, D.G., and THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- KAIER, A.N. (1895/6). Observations et expériences concernant des dénombrements représentatifs. *Bulletin of the International Statistical Institute*, 9, 176-183.
- KAIER, A.N. (1897). *The Representative Method of Statistical Surveys* (1976, English translation of the original Norwegian). Oslo: Central Bureau of Statistics of Norway.
- KAIER, A.N. (1905). Untitled speech with discussion. *Bulletin of the International Statistical Institute*, 14, 119-134.
- KISH, L. (1957). Confidence intervals for clustered samples. *American Sociological Review*, 22, 154-165.
- KISH, L., and FRANKEL, M.R. (1970). Balance repeated replication for standard errors. *Journal of the American Statistical Association*, 65, 1071-1094.
- KISH, L., and FRANKEL, M.R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society B*, 36, 1-37.
- KLEIN, L.R., and MORGAN, J.N. (1951). Results of alternative statistical treatments of sample survey data. *Journal of the American Statistical Association*, 46, 442-460.
- KONIJN, H.S. (1962). Regression analysis in sample surveys. *Journal of the American Statistical Association*, 57, 590-606.
- KRUSKAL, W., and MOSTELLER, F. (1980). Representative sampling, IV: the history of the concept in statistics 1895 - 1939. *International Statistical Review*, 48, 169-195.
- LANSING, J.B., and MORGAN, J.N. (1971). *Economic Survey Methods*. Ann Arbor: Survey Research Center.
- MAHALANOBIS, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, 109, 325-378.
- MANDEVILLE, J.P. (1946). Improvements in methods of census taking and survey analysis. *Journal of the Royal Statistical Society*, 109, 111-129.
- MCCARTHY, P.J. (1969). Pseudo-replication: half samples. *Review of the International Statistical Institute*, 37, 239-264.
- MURTHY, M.N. (1967). *Sampling Theory and Methods*. Calcutta: Statistical Publishing Society.
- NEYMAN, J. (1934). On the two different aspects of the representative method: stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-625.
- PFEFFERMANN, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61, 317-337.
- PORTER, R.D. (1973). On the use of survey sample weights in the linear model. *Annals of Economic and Social Measurement*, 2, 141-158.
- RAO, J.N.K., and BAYLESS, D.L. (1969). An empirical study of stabilities of estimators and variance estimators in unequal probability sampling of two units per stratum. *Journal of the American Statistical Association*, 64, 540-559.
- RAO, J.N.K., and SCOTT, A.J. (1981). The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76, 221-230.
- RAO, J.N.K., and SCOTT, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *Annals of Statistics*, 12, 46-60.
- RAO, J.N.K., and THOMAS, D.R. (1988). The analysis of cross-classification data from complex sample surveys. *Sociology Methodology*, 18, 213-269.
- RAO, J.N.K., and WU, C.F.J. (1987). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 321-241.

- SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- SCHUCANY, W.R., MINTON, P.D., and SHANNON, B.S. (1972). A survey of statistical packages. *Computing Surveys*, 4, 65-79.
- SEDRANSK, J. (1965). A double sampling scheme for analytical surveys. *Journal of the American Statistical Association*, 60, 985-1004.
- SHAH, B.V. (1978). SUDAAN: Survey data analysis software. *Proceedings of the Statistical Computing Section, American Statistical Association*.
- SHAH, B.V. (1984). Software for survey data analysis. *American Statistician*, 38, 68-69.
- SIMPSON, H.R. (1961). The analysis of survey data on an electronic computer. *Journal of the Royal Statistical Society, A*, 124, 219-226.
- SKINNER, C.J., HOLT, D., and SMITH, T.M.F. (1989). *Analysis of Complex Surveys*. New York: Wiley.
- STAFFORD, J.E., and ANDREWS, D.F. (1993). A symbolic algorithm for studying adjustments to the profile likelihood. *Biometrika*, 80, 715-730.
- STAFFORD, J.E., and BELLHOUSE, D.R. (1997). A computer algebra for sample survey theory. *Survey Methodology*, 23, 3-10.
- WILLCOX, W.F. (1926). The past and future developments of vital statistics in the United States I: John Shaw Billings and federal vital statistics. *Journal of the American Statistical Association*, 21, 257-266.
- WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.
- WOODRUFF, R.S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 66, 411-414.
- WORTON, D.A. (1998). *The Dominion Bureau of Statistics: A History of Canada's Central Statistical Office and Its Antecedents, 1841-1972*. Montreal and Kingston: McGill-Queen's University Press.
- YATES, F. (1960). *Sampling Methods for Censuses and Survey*, (3rd edition). London: Griffin.
- YATES, F. (1973). The analysis of surveys on computers – features of the Rothamsted Survey Program. *Applied Statistics*, 22, 161-171.
- YATES, F., and SIMPSON, H.R. (1960). A general program for the analysis of surveys. *Computer Journal*, 3, 136-140.

The Past is Prologue

BARBARA A. BAILAR¹

ABSTRACT

Mahalanobis provided an example of how to use statistics to enlighten and inform government policy makers. His pioneering work was used by the US Bureau of the Census to learn more about measurement errors in censuses and surveys. People have many misconceptions about censuses, among them who is to be counted and where. Errors in the census do occur, among them errors in coverage. Over the years, the US Bureau of the Census has developed statistical techniques, including sampling in the census, to increase accuracy and reduce response burden. A root-mean-square-error model was developed to estimate the joint effects of variance and bias in the census. The model is used in this paper to look at the joint effects of response variance, adjustment of the bias caused by the undercount, and the use of sampling for follow-up.

KEY WORDS: Censuses; Mahalanobis; Root-mean-square-error model; Sampling in the census.

1. INTRODUCTION

Perhaps it has always been so – that statistics, as a body of information, does not always support the actions that politicians want to take. In some countries, data from censuses are not made public, because knowledge is power. However, in our society, the power of statistics is used to inform us about needs for action, or how well we're doing as a country, or as the basis of comparison among groups. We are used to seeing and trusting statistics on an everyday basis, though most of us give little attention to how they are produced, by whom, and at what cost.

Over the last few decades, there have been many issues where statistics and politics have been in conflict. Employment and unemployment data are often used by politicians, especially in an election year. If the unemployment figures are low, the incumbents cite that figure and take the credit. If the employment figures show that many new jobs are being created, that number is cited. Either political party can use these data to make whatever political points seem salient. An attempt by the Nixon Administration to restrict access to these data led to new protections, such that the employment and unemployment data are released on the first Friday of every month by the Commissioner of the Bureau of Labor Statistics at a meeting of the Joint Economic Committee on Capitol Hill.

The definition of poverty is currently under discussion. When the poverty measure was invented by Molly Orshansky, there were not the large transfer payment systems that exist today. Because of income received or benefits paid, poverty today does not mean what poverty did 30 years ago. However, each political administration watches the poverty numbers very closely. These numbers were used by critics of the Reagan Administration to illustrate the growing burden of the poor in an administration that was alleged to be more interested in serving the rich. That Administration argued that by including medical

benefits and other transfer payments, the poor were better off than before.

Probability samples of the U.S. population are now used to study sexual behavior. Much of our information on sexual behavior goes back to Kinsey. The National Opinion Research Center (NORC) at the University of Chicago has conducted two large surveys of sexual behavior in the U.S. One of these, *Sex in America*, (Michael, Gagnon, Laumann and Kolata 1994) reported on a national sample of persons aged 18-59, and was not funded by the government. The second researched the sexual behavior of adolescents and, in both cases, federal funding for these studies was questioned because powerful constituencies did not want the subject matter to be examined. The second study was finally funded by the government.

Privacy issues abound. For example, there is broad concern about the confidentiality of individual medical records and the need for researchers to access them. Privacy issues for groups are less widely recognized. Certain groups may not want to report fully in a decennial census or survey because they do not want to attract attention. Though people who are in the country illegally are supposed to be included in the census, many of them fear that government authorities looking at block statistics could use the information to raid certain blocks.

My last example here of issues in which politics and statistics are having a disagreement, is the decennial census. For decades, an undercount in the census and its differential impact on minority populations has been well-documented. The Census Bureau has studied this issue for years and now has the statistical tools and methods to represent the uncounted individuals in the census totals. Yet this "adjustment" is opposed by many politicians because of an anticipated effect on the drawing of election district boundaries. However, the uses of the census extend far beyond apportionment and redistricting. The battle before the 2000 Census has been unusually intense.

¹ Barbara A. Bailar, National Opinion Research Center, 1155 East 60th Street, Chicago, Illinois, U.S.A.

Given these instances in which politics and statistics are confronting each other, it is useful to step back in time to review the contributions of Mahalanobis to the government of India. His methods were used successfully by the U.S. Census Bureau to learn much of what we know about errors in the census. I will review Mahalanobis' contributions, then return to a discussion of the census, the statistical tools currently used in the census, additional tools that could be used, and then conclude with a plea for Congress and the Census Bureau to follow the tradition of continuous improvement in the census through the use of statistical tools.

2. THE MAHALANOBIS LEGACY

Mahalanobis played an important role in the methodology we take for granted today. He was trained to teach physics, but became increasingly interested in statistical problems and then in building the Indian Statistical Institute. His work on the utilization of interpenetrated subsamples of the population was innovative, and gave great impetus to research on the effects of interviewers on survey and census statistics. He paid great attention to the need for pilot studies to test the implementation of survey techniques. As time went along, he enlarged his interests from sampling and surveys, in which he provided much needed information to the government, to planning and economic development. He was appointed Honorary Statistical Advisor to the Cabinet in January, 1949 and placed in charge of the Central Statistical Unit in the same year. The central role of statistics in government planning was, no doubt, due to the force of the man himself as well as his research findings. He saw the role of statistics as a system to serve the cause of planned development and envisioned a feedback arrangement between statistics and planning (Rudra 1996).

The particular contributions I wish to stress today are his major roles in sample surveys and in measuring error of all kinds – errors of observation, errors of measurement, sampling errors, copying errors, printing errors. Much of his early work on showing the variability in statistics caused by interviewers was in crop statistics (Mahalanobis 1950). He was one of the first to say, and then show, that the overall error in survey statistics was not just sampling variance but also the variance arising from the human element. One way to study such errors was by the use of interpenetrated subsamples. In the words of Mahalanobis,

“When two (or more) samples are drawn from the same population and covered according to the same survey design, the results based on the different samples are equally valid, even though they are derived by different operational units; and divergences between the different sets of estimates supply directly some idea of the margin of uncertainty.” (Mahalanobis and Lahiri 1961)

Mahalanobis demonstrated that statistics based on samples were at least comparable to, and often more accurate than statistics based on a census, in the 1940's, when sampling was still not fully accepted. He believed, as many of us now do, that samples can be better controlled than can a census. He stated (Mahalanobis and Lahiri 1961) that the magnitude of discrepancies found in a census of jute production made it appear that a census may not provide accurate estimates for small areas. The random component of the non-sampling error may add enough error that results for a large area may be no different from those obtained by a sample survey. What holds for a large area does not naturally follow for small areas.

The U.S. Census Bureau used Mahalanobis' techniques to learn more about the underlying variability of census numbers.

3. WHAT DO PEOPLE THINK A CENSUS IS

To most people, taking a census means that enumerators go out and count everyone. There are three things that people seem to think about censuses. One is that everyone is counted. A second is that an enumerator sees everyone. A third is that the census is without error. Let's look at these one by one.

Often, everyone is not supposed to be counted in a national census, and who should be counted varies from country to country, and over time within a country. For example, military personnel and their families located outside of the country could be counted or not. Civilian aliens temporarily in the country as seasonal workers could be counted or not. From these illustrations one can see that a primary necessity in census-taking is defining the scope of the census.

So, by definition, certain groups of people are not to be counted in the census. This is by design of the Census Bureau. Other people make individual or family decisions not to be counted in the census. In earlier times, some families did not report children who suffered from some diseases or retardation. Some people who have had unfortunate episodes with the legal system may decide not to be counted. These may be people who are in the country illegally, those who are hiding from law enforcement, and those who fear, for whatever reason, the consequences of being counted. In 1990, there were people who said they would not be counted because they thought the census was too intrusive.

Finally, there are people missed, not by design but by accident. Perhaps they lived in buildings that were missed, perhaps they lived on the street and were missed. Perhaps they were away during the census period. During 1998 there were many reports of how much harder it was to survey people who live in gated communities. It may be that some of these people are missed because of the overzealousness of the community guards. In some communities good

maps are unavailable or not updated, so groups of people may be missed.

In any case, not everyone is counted in a census and never was.

The second myth to be refuted is that an enumerator sees everyone and knows who should be in the census or not. This never happened, even in the early censuses in the U.S., when U.S. Marshals took the census and the country was much smaller. In fact, early censuses were of households, not of individuals. This means that there were no questions asked of individuals but instead there was interest in how many people were in the household, how many were men and how many were women, how many were in different age groups, and so forth. The totals were posted in public places. Starting in 1880 the canvasser method of taking a census, where enumerators went from door to door, came into being. It is this kind of census that made some believe that an enumerator saw everyone. However, a single household member usually responded for the whole family. The enumerator did not see those who were sick, those at work, those who were away temporarily, or those who were, for some reason or another, not in the room when the enumerator visited.

Though the enumerator-type census was an improvement over one taken by the marshals, research using interpenetrated subsamples showed that census enumerators still added a considerable amount of error to the census statistics. The enumerators were influenced by their own expectations and by responses of others in their enumeration district. Also, some did not understand the instructions and reported things incorrectly. An experiment in the 1950 census showed that enumerators added considerable variability to the census statistics (Hanson and Marks 1958). Indeed, the statistics gathered from a census had the same level of variability, due to enumerators, as a 25-percent sample. This is the main reason the Census Bureau turned to the use of self enumeration in the 1960 census and progressively expanded it in later censuses. Now, if a household receives the census form by mail, fills it out, and sends it in, and no errors require resolution, no enumerator will call at the household.

The third myth is that census taking occurs without error. No one who now works on censuses or surveys believes that, but other people do. The Census Bureau encourages that belief by publishing data down to the last digit. For example, the population of the United States in 1990 was reported and published as 248,718,301 in the *Statistical Abstract*.

Even some of those who have worked closely with a census cannot see it as a statistical process that carries with it a certain amount of error. Because the error is not routinely quantified and published along with the census numbers, some cannot believe the error exists. Some persons working in the Population Division of the U.S. Census Bureau in the 1940's and 50's believed that the census was the best way to learn about any subject, and that

sample surveys were inferior. Repeated demonstrations of accuracy in survey results and of bias in census data did not change their minds.

Anyone who comes into regular contact with the census now knows that there is error in the data. First, though sampling variance cannot occur for items collected on a 100-percent basis, there may still be substantial response variance introduced by effects of enumerators, respondents, and coders on census data. Second, bias affects responses to many census questions even when a person is correctly counted. Bias also affects counts when enumerators do not count everyone. The Census Bureau conducts an evaluation program as part of every census, documents the amount of error, and uses those data to attempt to improve the next census.

Large groups of people are affected by census error. The undercounting bias affects minority populations and children at a much higher rate than other populations (Edmonston and Schultze 1993). Thus, communities that are largely African-American, Hispanic, or American Indian are underrepresented in distributions of potential power and money, while those statistics that are based on children under 10 are subject to a large error.

Over the years, the Census Bureau has reported numerous studies looking at the balance between cost and accuracy. One mentioned before is the use of self-enumeration. At smaller levels of population, the effect of response variance, primarily caused by interviewers, was very high. Just as with sampling error, as the size of the area increased, and the number of enumerators who collected the data increased, the effect lessened. When the mail return rate was close to 80 percent, the response variance decreased to about one-quarter of that of a 25-percent sample (Bailar 1969).

Thus, commonly held images of the census are not always true. Also, the census is not always the same. The Census Bureau has made many changes in census taking since the first census in 1790. The number of questions, the kinds of questions, who is counted and where, who does the counting, how people are assigned to a geographic domain, how missing characteristics are handled, and the gradual increase of asking most questions of a sample have changed over the years. The next section shows how the use of statistical tools has changed the census in this century.

4. DEVELOPMENT OF STATISTICAL TOOLS IN A CENSUS

Two elements have changed the methods of the U.S. Decennial Census considerably since 1940: the use of computers; and the use of statistical techniques. At times, the two elements have complemented each other, for example in the fast processing for imputation of missing data using a "hot deck" procedure. While computers have profoundly affected the census, the remainder of this discussion will focus on the statistical methodology.

One of the major advances starting in 1940 has been the use of sampling in the census. In 1940, as documented by Waksberg and Hanson (1965), there were three major uses of sampling. One was for the collection of data deemed supplementary to the main census questions. Questions such as mother tongue, veteran status, and fertility were asked of a 5-percent sample. A second use was for certain analytic studies requiring clerical transcription and coding. To avoid a long timespan for the transcription and coding to take place, a sample of census questionnaires was selected and the transcription and coding occurred only for them. A third use was for the verification of large-scale clerical operations such as editing, coding, key-punching, and so forth. Prior to 1940, all verification was on a 100-percent basis.

To describe the next leap forward, Waksberg and Hanson said:

“A major step forward in the use of sampling in census work took place in the 1950 Census of Population and Housing. This grew out of a profound change in attitude regarding the role of sampling. Whereas in 1940 sampling had been considered applicable only for items of supplementary and secondary interest, in 1950 the entire range of census activities was examined to determine, on a logical basis, where complete counts were necessary and where samples could provide adequate information.”

The increased use of sampling for population characteristics, for sample tabulations, and for verification was successful and evaluations showed that, even with the addition of sampling error, overall error was less than if earlier techniques had been used with no sampling. This was a reinforcement of the lesson learned earlier by Mahalanobis.

During the 1950 Census, the Bureau did a great deal of research to learn the effect of response biases and response variances on census data. Waksberg and Hanson declared that it was misleading to assume that the census, without sampling, was without error. In 1950, an experiment was conducted to estimate the effect of census enumerators on census data. By using the method of interpenetrated subsamples introduced by Mahalanobis, pairs of adjacent census areas were merged and assignments to the enumerators were randomized. Since the assignments were over the same area, differences between enumerators did not reflect differences in the type of area. The main finding of the study was that a full census in which enumerators went door to door to collect the census information had response variability that made the census the equivalent of a 25-percent sample (Hanson and Marks 1958). Using that result, as well as studies of biases in various census items, Waksberg and Hanson formulated a model in which census results were subject to a relative response bias of 6 percent and a response variance equal to the sampling variance of a 25-percent household sample. They used this model to generate Table 1 which shows the magnitude of total error in census data with and without sampling.

The authors point out that for a characteristic describing 500 individuals in an area of 2,500 people, the increase in the total root mean square error arising from sampling variability is only about 25%. For larger areas and larger cells, the additional error due to sampling is even smaller.

These data were studied carefully before the decision to increase the use of sampling in the 1960 Census. In practice, sampling made even greater gains than those anticipated by the model. The authors state “Thus for a great many published statistics, the reliability was better with the use of sampling than would have been possible otherwise.” (Waksberg and Hanson 1965.)

Table 1
Expected Root Mean Square Error (RMSE) of Estimated Cell Frequencies for Individual Items Based on a Complete Census and On a 25-Percent Sample of Households

Area of 2,500 Population having RMSE based on			Area of 10,000 Population having RMSE based on			Area of 50,000 Population having RMSE based on		
Cell Frequency	Complete Census	25-percent Sample	Cell Frequency	Complete Census	25-percent Sample	Cell Frequency	Complete Census	25-percent Sample
12	7	10	50	1	20	250	34	46
50	14	19	200	30	40	1,000	85	105
125	22	31	500	52	67	2,500	180	200
500	49	62	2,000	140	160	10,000	620	650
1,250	89	102	5,000	320	330	25,000	1,520	1,530

Note 1: Computations assume a relative response bias of 6 percent and response variance equal to the sampling variance for a 25-percent sample.

Note 2: The accuracy of the results (cell frequencies) is measured by a certain kind of average of the actual errors that would occur, the root mean square error (RMSE). A useful working rule would be to assume that approximately two-thirds of all results of a census or a sample would differ from their true cell frequencies by no more than their RMSE's.

A large-scale evaluation and research program of the Decennial Census program began in the 1950's and is now an integral part of the Census. Part of the program tests new methods for possible use in the following census and part of it focuses on the evaluation of the current census. It was as part of this program that the Bureau started measuring the undercount in the census. It was also this program in which the response variance due to enumerators was measured before and after the advent of self-enumeration. (In 1960, after self-enumeration was introduced, the response variance decreased to 1/4 of the 1950 level.) Since mail-back rates have decreased substantially since 1980, that variance may have increased again, perhaps substantially.

Other studies included research on alternative ways to measure the undercount, record checks to measure the accuracy of census data, and a study of using the Post Office not only to deliver census questionnaires but to notify the Census about missed addresses and duplicate forms.

Sampling is now used extensively to control the quality of the large-scale clerical tasks associated with the census. In past censuses, verification was usually dependent, in which the verifier reviewed the coder's work and determined whether the correct codes had been assigned. The Bureau planted errors and found that dependent verification missed as many as half of the errors. This and other research caused the Bureau to develop independent verification, in which records are assigned to three coders who do not see each other's work. A "majority rule" is used to determine the best code, and statistics about such errors are used to improve the process and to identify substandard performance.

Imputation was also a necessary tool developed for use in the census. To keep within time and budget parameters, the Bureau developed a "hot-deck" imputation system, based on the assumption that people who live in proximity are likely to resemble each other for many characteristics such as educational attainment and income. Another kind of imputation was also used in 1970, 1980, and 1990 to deal with a small, residual set of addresses left on the mailing list with no information about whether or not they were occupied. No one answered the door, nor did neighbors know if anyone lived there. Thus, based on a model that assumed a high correlation between the characteristics of neighboring households, the Bureau imputed occupancy or vacancy status, and to those imputed as occupied, a number of people were imputed. In 1980, only 762,000 persons were imputed, about .003 of the total census count, but they were not spread evenly over all the States. As a result of the imputation, Indiana lost a Congressional seat to Florida. However, it should be acknowledged that doing nothing about the unclassified units would have been equivalent to imputing them all as vacant. There was information available that showed that over half of these units could be expected to be occupied so the data based on imputation were more accurate than data based on counts alone with no imputation.

5. ADDITIONAL USES OF STATISTICAL TOOLS

Statistical tools can be used to correct the census for the undercount. The Waksberg-Hanson root mean-square error model estimates the amount of error in the census assuming a relative response bias in the overall census of 2 percent. (The 1990 estimate was 1.6 percent.) Also assume a response variance in both the adjusted and unadjusted census equal to one-fourth the sampling variance of a 25-percent sample. That estimate may now be too low since decreasing mail-back rates have driven enumerator variances higher. However, to be on the conservative side, we shall use the 1960 and 1970 measurements.

The model is the simple mean-square error model used frequently by the Census Bureau.

$$MSE(T) = \text{Var}(T) + B_T^2$$

Assume T is a cell size or a size of interest in the census in an area where N is the population size. $T = NP$ where P is the proportion of the population having a certain characteristic. B is the bias in the census count. So, for example, in an area of 2,500 people, one might be interested in knowing the number of children under 10 years of age. $N = 2,500$ and $T = NP$.

Now the variance of an estimated proportion, p is:

$$V(p) = \frac{N-n}{N-1} \cdot \frac{1}{n} \cdot PQ$$

If we have a 25 percent sample, this reduces to

$$V(p) = \frac{3}{4} \cdot \frac{1}{n} \cdot PQ = \frac{3PQ}{N}$$

$$V(T) = V(Np) = N^2 V(p) = 3NPQ \\ = 3TQ$$

Relative bias = (.02) so Bias = .02T

Now we are dealing with a census, so there is no sampling variance, but the response variance is equal to 1/4 of what the sampling variance would be. So

$$MSE(T) = (.02T)^2 + (.25)(3)TQ$$

and

$$RMSE(T) = \sqrt{(.02T)^2 + (.25)(3)TQ}$$

This formula has been used as the basis of the calculations in Table 2. For an unadjusted census, $RMSE(T)$ would have both the bias and variance components. For an adjusted census, the relative bias is zero, so only the response variance term remains. However, this analysis presumes that the adjustment factors themselves are free from any kind of variance and bias, and that the same adjustment factors can be uniformly applied within the demographic groups.

For example, Table 2 shows that for a total of 500 in an area of 2,500, the RMSE for an unadjusted census is 20 while the RMSE for an adjusted census is 17. For the unadjusted census, the contribution from the bias term is small, $[(.02)(500)]^2=100$. The contribution from the response variance is $(.25)(3)(500)(.8)=300$. So $RMSE = \sqrt{400}=20$. For the adjusted census, the bias term, 100, is removed, so the $RMSE=\sqrt{300}=17$. However, if one considers that the estimated bias term has both variance and bias, there may be little difference between the adjusted and unadjusted results for a small area. As the total, T , gets larger, the bias term is more dominant, and the adjustment removes more error.

Table 2 shows that for a small area of 2,500 persons there is no gain for small totals, but a gain of 43 percent in accuracy for a large total of 1,500 persons. In a somewhat larger area of 10,000 persons, there is little reduction in error until a total of 1,000 is of interest, where there is a gain in accuracy of 21 percent and for a large total of 5,000, there is a gain of 61 percent. Thus, if we were talking about the number of men or women in an area of 10,000, a total that might be expected to be around half the population, there would be a large gain in the accuracy of the total from using adjusted census figures. For an area of 50,000 the bias term dominates the mean square error, even at smaller totals such as 1,000. Here the gain is 21 percent, which grows to 81 percent for a very large total of 25,000.

This illustration shows is that adjusting the census does not add to the error of the census, even for small areas and small cells, if one assumes that the bias term is measured without error. For smaller area sizes and smaller cells, the response variance dominates the mean square error, but the total error is never less than the response variance. When

the census is adjusted, the bias term goes to zero, and the gains in accuracy are dramatic.

One virtue of this model is that it was developed by the Census Bureau long before the current debate on adjustment grew heated. It was used to disabuse people of the idea that the census cells have no error. It was used successfully to show critics that having most of the census questions answered by only a sample would not hurt the data unduly. Such a tried and true census model now shows the real value of adjustment.

Table 2 used the relative response bias of 2 percent based on the 1990 Census overall estimate of the undercount of 1.6 percent. However, since the undercount hits minority populations harder, let's look at a comparison of an adjusted and unadjusted census in which the relative bias is 4 percent. (The 1990 estimates of the undercount were 4.4 percent for African-Americans, 4.5 percent for American Indians, 5.0 percent for Hispanics, and 2.3 percent for Asians.)

Table 3 shows the RMSE for minority communities for the sizes 2,500, 10,000 and 50,000. Though the RMSE's for the adjusted census stay the same, since the bias has been removed, the unadjusted RMSE's are considerably larger. The gains in accuracy from an adjustment are much larger in minority communities, as one would expect. For example, as shown above, the error in the number of males in a non-minority community of 10,000 would be about 109 unadjusted and 43 adjusted. In a minority community, the errors are 205 and 43 respectively. In a larger area of 50,000 the improvement is dramatic even for a small cell of 1,000.

Now, suppose we repeal the 1976 law that specifies that there shall be no sampling for the apportionment numbers. Think about a census in which, after a certain date, the housing units not returning census forms are sampled.

Table 2
Expected Root Mean Square Error (RMSE) of Estimated Cell Frequencies for Population Estimates Based on a Census with No Adjustment for Undercount and with Adjustment

Area of 2,500 Population having RMSE based on			Area of 10,000 Population having RMSE based on			Area of 50,000 Population having RMSE based on		
Cell Frequency	Unadjusted Census	Adjusted Census	Cell Frequency	Unadjusted Census	Adjusted Census	Cell Frequency	Unadjusted Census	Adjusted Census
15	3	3	50	6	6	250	15	14
50	6	6	100	9	9	500	22	19
100	9	8	200	13	12	1,000	34	27
500	20	17	500	21	19	2,500	65	42
750	25	20	1,000	33	26	5,000	116	58
1,000	29	21	2,000	53	35	10,000	214	77
1,500	37	21	5,000	109	43	25,000	509	97

Note: Computations assume a relative response bias of 2 percent in the unadjusted census and 0 percent in the adjusted census. There is a response variance in both the adjusted and unadjusted census equal to 1/4 the sampling variance of a 25 percent sample.

Table 3

Expected Root Mean Square Error (RMSE) of Estimated Cell Frequencies for Population Estimates in African-American, American Indian, and Hispanic Communities Based on a Census with No Adjustment for Undercount and with Adjustment

Area of 2,500 Population having RMSE based on			Area of 10,000 Population having RMSE based on			Area of 50,000 Population having RMSE based on		
Cell Frequency	Unadjusted Census	Adjusted Census	Cell Frequency	Unadjusted Census	Adjusted Census	Cell Frequency	Unadjusted Census	Adjusted Census
15	3	3	50	6	6	250	17	14
50	6	6	100	9	9	500	28	19
100	9	8	200	15	12	1,000	48	27
500	26	17	500	21	19	2,500	109	42
750	31	20	1,000	48	26	5,000	208	58
1,000	45	21	2,000	87	35	10,000	407	77
1,500	64	21	5,000	205	43	25,000	1,004	97

Note: Computations assume a relative response bias of 4 percent in the adjusted census and 0 percent in the adjusted census. The response variance in both the adjusted and unadjusted census equal to 1/4 the sampling variance of a 25 percent sample.

In this model, there are two components of variance, the response variance and the sampling variance. The sampling variance is based only on the nonresponse universe.

Let R be the nonresponse rate, and M the population of nonresponse households. Then $M = RN$. The total for which we are trying to estimate the sampling variance is $S = PM$. The relationship between S , the sampled part of the total, and T , the total, is through R . $S = PM = P(RN) = RT$.

So the sampling variance = $3MPQ = 3PQRN$, assuming a 25-percent sample of the nonrespondents. This sampling rate could easily be changed for a larger rate, but for purposes of illustration, it suffices.

In Table 4, there are three contributors to the RMSE. Two of them are the terms we saw in the earlier description when sampling of the non-mail returns was not a consideration. Now we have a third term, expressing the sampling variance arising from the sample of non-mail returns. In an adjustment, only the bias term goes to zero, while the two variance terms remain. Each of the variance terms gets smaller as the cell size gets larger, but they do not vanish.

Table 4 shows the RMSE's for a census with no sampling of non-mail return households, with and without adjustment, for a 25 percent sample of non-mail return households when only half of the population mails them back and when 70 percent mail them back for the three sizes of area we have looked at before: 2,500 population, 10,000 population, and 50,000. The no sampling case is what we will have in the 2000 Census because the use of sampling for follow-up is prohibited. Look first at Section A for a population of 2,500. Where there is no sampling of non-mail return households, we see the numbers from Table 2. When half of the population mails back the census form, and the remaining half is sampled, the variance component keeps the adjusted and unadjusted RMSE's very close together. At maximum, there is a 20 percent reduction in

error. There is somewhat more gain when the mailback rate is .70 and only 30 percent of the remaining population is sampled. The maximum gain in this case is 28 percent.

Small areas, such as those of 2,500 may be greatly affected by sampling, especially at a 25-percent rate if the mailback rate is low. Whether a decrease in accuracy is acceptable depends on the uses for the data. Since providing small area data is an important objective of the census, it may be that there would need to be a much larger sampling rate, if not complete follow-up for small areas. The Census Bureau has done this before with some characteristics, such as income, so that there would be less variability in the income data for areas of 2,500 or fewer persons. Following that same principle, it could be specified that there would be no sample follow-up in places of 2,500 persons or fewer, and variable follow-up rates depending on place size. Another strategy would be to use the information abundantly available about coverage error and to specify larger samples in places that have characteristics highly correlated with the undercount.

For areas of 10,000 population, we see a definite improvement from the adjustment for the bias, but the adjusted numbers with sampling are still considerably larger than the adjusted figures without sampling. However, if there is no adjustment, the sampling adds to the RMSE, but the unadjusted numbers are not much different. There is a 15 percent increase in the RMSE when only half the population returns the census form and an increase of 9 percent when 70 percent return it.

Finally, when we look at an area of 50,000 we see that the bias dominates the RMSE for all but the smallest cell sizes. When the total we are trying to estimate is 5,000 or larger, sampling adds to the RMSE, but an adjustment, with sampling, is still superior to unadjusted numbers with no sampling.

Table 5 is similar, but geared to a predominantly minority population. As in Table 3, the relative bias is 4 percent, reflecting an average undercount rate for minorities. In this table, the RMSE's for unadjusted totals are much more similar, even for smaller areas, because of the larger effect of the bias term on the RMSE.

The results for areas of 50,000, which exist in most large cities, show the devastating effects of not adjusting for the large minority undercount. The sampling variance for the

larger totals has practically no effect on the RMSE, but the improvement from adjustment for all cases, sampling or no sampling is 83 percent or higher. The added error because of sampling is negligible.

Unfortunately, in many minority communities, low mail-return rates and undercounting occur together. Such communities have a 50 percent mail return rate or lower. It may be that the sample size will need to be increased in these areas.

Table 4

Expected Root Mean Square Error (RMSE) of Estimated Cell Frequencies for Population Estimates Based on an Unadjusted Census, an Adjusted Census, and on a 25 Percent Sample of Non-Mail Return Households

A. Area of 2,500 population having RMSE based on

Cell Frequency	No sampling of non-mail return HH's		25% sample and .50 mailback rate		25% sample, and .70 mailback rate	
	Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
15	3	3	6	6	5	5
50	6	6	11	11	9	9
100	9	8	15	15	13	13
500	20	17	32	30	28	26
750	25	20	38	34	33	29
1,000	29	21	42	37	37	31
1,500	37	21	47	37	43	31

B. Area of 10,000 population having RMSE based on

Cell Frequency	No sampling of non-mail return HH's		25% sample and .50 mailback rate		25% sample, and .70 mailback rate	
	Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
50	6	6	11	11	9	9
100	9	9	15	15	13	13
200	13	12	21	21	18	18
500	21	19	34	33	30	28
1,000	33	26	49	45	43	39
2,000	53	35	72	60	65	51
5,000	109	43	125	75	119	64

C. Area of 50,000 population having RMSE based on

Cell Frequency	No sampling of non-mail return HH's		25% sample and .50 mailback rate		25% sample, and .70 mailback rate	
	Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
250	15	14	24	24	21	20
500	22	19	35	33	30	29
1,000	34	27	51	47	45	40
2,500	65	42	89	73	80	63
5,000	116	58	142	101	132	86
10,000	214	77	241	134	231	115
25,000	509	97	527	168	520	144

Table 5

Expected Root Mean Square Error (RMSE) of Estimated Cell Frequencies for Population Estimates in African-American, American Indian, and Hispanic Communities Based on an Unadjusted Census, an Adjusted Census, and on a 25 Percent Sample of Non-Mail Return Households

A. Area of 2,500 population having RMSE based on

Cell Frequency	No sampling of non-mail return HH's		25% sample, and .50 mailback rate		25% sample, and .70 mailback rate	
	Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
15	3	3	6	6	5	5
50	6	6	11	11	9	9
100	9	8	15	15	13	13
500	26	17	36	30	33	26
750	31	20	46	34	42	29
1,000	45	21	54	37	51	31
1,500	64	21	70	37	68	31

B. Area of 10,000 population having RMSE based on

Cell Frequency	No sampling of non-mail return HH's		25% sample, and .50 mailback rate		25% sample, and .70 mailback rate	
	Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
50	6	6	11	11	9	9
100	9	9	15	15	13	13
200	15	12	22	21	18	18
500	21	19	38	33	34	28
1,000	48	26	60	45	56	39
2,000	87	35	100	60	95	51
5,000	205	43	214	75	210	64

C. Area of 50,000 population having RMSE based on

Cell Frequency	No sampling of non-mail return HH's		25% sample, and .50 mailback rate		25% sample, and .70 mailback rate	
	Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
250	17	14	26	23	23	21
500	28	19	39	33	35	29
1,000	48	27	62	47	57	40
2,500	109	42	124	73	118	63
5,000	208	58	224	101	218	86
10,000	407	77	422	134	416	115
25,000	1,004	97	1,014	168	1,010	144

6. CONCLUSION

It has been a tradition for the Census Bureau in the latter half of this century to use statistical techniques, where possible, to make the Decennial Census more accurate and less costly. Using the techniques historically used by the Census Bureau, namely a mean-square error model, one can see that adjustment does improve census totals, even for small areas, when one assumes even a minimal level of response variance. One can also see the need for precaution if sampling is to be used for follow-up. It may be that there should be no sampling in places of 2,500 or fewer people, just as there is no sampling for certain population characteristics in these small places.

In looking at the current census controversy, it is good to remember the spirit of Mahalanobis. Not only did his ingenious use of interpenetrated subsamples give us the ability to estimate the response variance in census statistics, but his insistence that sampling and statistics should be used to solve practical problems has been the hallmark of the U.S. Census. Some of the most fundamental practical problems are those faced by the government and Mahalanobis allocated statistical resources for the solving of these problems. Likewise, the U.S. Census Bureau has a long and rich history of offering practical, cost-efficient solutions to thorny census problems.

REFERENCES

- BAILAR, B.A. (1969). Evaluation and Research Program of the U.S. Censuses of Population and Housing, 1960: The Effect of Interviewers and Crew Leaders. Series ER 60 No. 7. Washington, DC: U.S. Bureau of the Census.
- EDMONSTON, B., and SCHULTZE, C. (1993). *Modernizing the U.S. Census*. Washington, DC: National Academy Press, 34-35.
- HANSON, R.H., and MARKS, E.S. (1958). The influence of the interviewer on the accuracy of survey results. *Journal of the American Statistical Association*, 53, 639-655.
- MAHALANOBIS, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, 325-378.
- MAHALANOBIS, P.C. (1950). Why Statistics? *Sankhyā*, 10, 195-228.
- MAHALANOBIS, P.C., and LAHIRI, D.B. (1961). Analysis of errors in censuses and surveys with special reference to experience in India. *Bulletin of the International Statistical Institute*, 38, 2, 401-433 (reprinted in *Sankhyā*, 23, 325-358).
- MICHAEL, R.T., GAGNON, J.H., LAUMANN, E.O., and KOLATA, G. (1994). *Sex in America, A Definitive Survey*. New York: Little Brown and Co.
- RUDRA, A. (1996). *Prasanta Chandra Mahalanobis: A Biography*. New York: Oxford University Press.
- WAKSBERG, J., and HANSON, R. (1965). Sampling Applications in Censuses of Population and Housing. U.S. Bureau of the Census, Technical Paper No. 13.

Estimation of Census Adjustment Factors

C.T. ISAKI, J.H. TSAY and W.A. FULLER¹

ABSTRACT

A components-of-variance approach and an estimated covariance error structure were used in constructing predictors of adjustment factors for the 1990 Decennial Census. The variability of the estimated covariance matrix is the suspected cause of certain anomalies that appeared in the regression estimation and in the estimated adjustment factors. We investigate alternative prediction methods and propose a procedure that is less influenced by variability in the estimated covariance matrix. The proposed methodology is applied to a data set composed of 336 adjustment factors from the 1990 Post Enumeration Survey.

KEY WORDS: Components-of-variance; Small area estimation; Undercount; Decennial Census; Smoothing.

1. INTRODUCTION

While the objective of a population census is to record data for all individuals, it has long been recognized that this goal is not achieved in practice. Post enumeration studies associated with the U.S. Census of 1970 and 1980 suggested that the coverage rate was different for different demographic groups. See U.S. Bureau of the Census (1988).

In 1990, a post enumeration survey (PES), using dual system (or capture-recapture) estimation, was used to produce estimates for 1392 subdivisions of the total population of the United States at the time of the 1990 Census. The PES sample contained approximately 377,000 persons in about 5200 sample blocks. Sample persons were divided into post-strata defined by geographic divisions of the country, tenure, size-of-place, race, sex, and age, where the two tenure classes are owners and renters of homes, and size-of-place is a measure of urbanization. The subdivisions were called poststrata. The ratio of the PES estimate to the Census total, called the adjustment factor, was produced for each poststratum. An adjustment factor greater than one is associated with an estimated undercount and a factor less than one is associated with an estimated overcount.

Because relatively large sampling variances were anticipated for individual ratios, a smoothing technique based on components-of-variance and a regression model was used to create the final estimated adjustment factors. The elements of the error covariance matrix used in the prediction model were estimated with a jackknife algorithm, see Fay (1990).

The explanatory variables in the regression model were chosen using a best subsets selection algorithm. Some explanatory variables were forced into the model. For example, in the Midwest region, the ten explanatory variables forced into the model were Black, Hispanic, renter, age group 0-9, age group 10-19, age group 20-29, age group 30-44, age group 45-64, male 10-19 and male 20-64. Most

variables were indicator variables, but some were proportions. For example, a variable "percent Black" was used when Black and Hispanic were grouped into a single poststratum. Nine other variables were selected for inclusion in the model based on a best subsets regression algorithm. The variables included mail return rate, substitution rate, type-of-place and six race-by-age and race-by-tenure interaction variables. The mail return rate is the fraction of Census questionnaires returned from the mail distribution, the substitution rate is the fraction of Census households that were entirely replaced with responding households.

The smoothing technique was applied to poststrata ratios by regions of the country. The adjustment factors were designed to be applied to Census counts in the appropriate poststrata to create population estimates adjusted for undercount or overcount. Hogan (1992) contains an overview of the PES. Isaki, Huang and Tsay (1991) provide a detailed description of the results of the smoothing of the poststratum ratios.

Fay (1992) in a manuscript discussing the adjustment factors constructed from the 1990 PES, identified some disturbing results. He noted that some of the estimated regression coefficients in the model differed considerably depending on the form of the estimated covariance matrix used to construct the estimated generalized least squares estimator. Fay conjectured that large differences in coefficients could arise because of an unstable estimator of the error covariance matrix. Although the estimated error variances were smoothed, it was felt that estimated variances of linear combinations might still have large variances. He felt that the estimated variances had large variances because the direct estimates for many blocks were zero.

The Secretary of Commerce ultimately decided to use the unadjusted counts in the Decennial Census. The possible use of adjusted counts for other purposes, such as the Bureau's postcensal estimation program, was left for additional study.

¹ C.T. Isaki and J.H. Tsay, Statistical Research Division, Bureau of the Census, Washington, D.C. 20233, U.S.A.; W.A. Fuller, Department of Statistics, Iowa State University, Ames, IA 50010, U.S.A.

We explore alternative smoothed estimators for the adjustment factors, focusing on the effect of estimating the covariance matrix of the vector of the estimated adjustment factors. In the empirical part of our study, we construct estimates based on the 1990 Census data.

2. SMOOTHING MODEL

The model chosen for the construction of predictors is the multivariate components-of-variance model. Closely related models that lead to smoothed estimators for a set of unknowns, have been studied by a number of authors. Fay and Herriot (1979) suggested the use of the model in a small area estimation procedure. Battese, Harter and Fuller (1988) applied the components-of-variance model to crop area estimation. Ericksen and Kadane (1985), Cressie (1992), and Ericksen, Kadane and Tukey (1989) suggested smoothing procedures for census adjustment. Singh, Gambino and Mantel (1994) discuss a range of small area procedures. Efron and Morris (1972) and Morris (1983) contain good discussions of some of the basic theory. Kackar and Harville (1984), Peixoto and Harville (1986), Fay (1987), Fuller and Harter (1987), Hulting and Harville (1991), Ghosh (1992), and Prasad and Rao (1990) discuss estimation and variance estimation for such procedures. Ghosh and Rao (1994) is a review article.

Under the multivariate components-of-variance model, the vector of true values to be predicted is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{w}, \quad (1)$$

where \mathbf{y} is an n -dimensional column vector, \mathbf{X} is an $n \times k$ matrix of observable characteristics, \mathbf{w} is an n -dimensional column vector of random effects and $\boldsymbol{\beta}$ is a k -dimensional unknown column vector. The vector \mathbf{Y} is observed, where

$$\mathbf{Y} = \mathbf{y} + \mathbf{e}, \quad (2)$$

\mathbf{Y} is an n -dimensional column vector and \mathbf{e} is the n -dimensional column vector of estimation errors. In our application \mathbf{Y} is the vector of estimated adjustment factors. It is assumed that

$$(\mathbf{w}', \mathbf{e}')' \sim \left(\mathbf{0}, \text{block diag} \left\{ \mathbf{I}\sigma^2, \boldsymbol{\Sigma}_{ee} \right\} \right), \quad (3)$$

where $\boldsymbol{\Sigma}_{ee}$ is the covariance matrix of the estimation errors, and σ^2 is the unknown variance of the random effects.

A class of predictors of \mathbf{y} is defined by

$$\tilde{\mathbf{y}} = \mathbf{X}\mathbf{B} + \mathbf{G}'(\mathbf{Y} - \mathbf{X}\mathbf{B}), \quad (4)$$

where \mathbf{B} is a k -dimensional vector and \mathbf{G} is an $n \times n$ matrix. Under model (1) with

$$(\mathbf{w}', \mathbf{e}')' \sim N\left(\mathbf{0}, \text{block diag} \left\{ \mathbf{I}\sigma^2, \boldsymbol{\Sigma}_{ee} \right\} \right), \quad (5)$$

the conditional expected value of \mathbf{y} given \mathbf{Y} is

$$E\{\mathbf{y} | \mathbf{Y}\} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}'_{zz}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}), \quad (6)$$

where $\mathbf{G}_{zz} = \boldsymbol{\Sigma}_{zz}^{-1}\sigma^2$ and $\boldsymbol{\Sigma}_{zz} = \mathbf{I}\sigma^2 + \boldsymbol{\Sigma}_{ee}$ is the $n \times n$ covariance matrix of $\mathbf{z} = \mathbf{w} + \mathbf{e}$. Under the normal distribution model defined by (1), (2), and (5) and with the parameters σ^2 , $\boldsymbol{\Sigma}_{ee}$, $\boldsymbol{\beta}$ known, the minimum mean square error predictor of \mathbf{y} is given by the right side of equation (6).

Generally, some of the parameters are unknown. Consider first the case in which $\boldsymbol{\beta}$ is unknown. Let $\hat{\boldsymbol{\beta}}$ be an estimator of $\boldsymbol{\beta}$, where

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{M}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}^{-1}\mathbf{Y}, \quad (7)$$

and \mathbf{M} is an $n \times n$ matrix. If \mathbf{M} is fixed

$$\begin{aligned} \tilde{\mathbf{y}} - \mathbf{y} &= (\mathbf{I} - \mathbf{G})\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - (\mathbf{I} - \mathbf{G})\mathbf{w} + \mathbf{G}\mathbf{e} \\ &= (\mathbf{K} - \mathbf{I})\mathbf{w} + \mathbf{K}\mathbf{e}, \end{aligned}$$

where $\mathbf{K} = (\mathbf{I} - \mathbf{G}')\mathbf{X}(\mathbf{X}'\mathbf{M}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}^{-1} + \mathbf{G}'$. Thus, if \mathbf{M} and \mathbf{G} are fixed,

$$\mathbf{V}\{\tilde{\mathbf{y}} - \mathbf{y}\} = (\mathbf{K} - \mathbf{I})(\mathbf{K} - \mathbf{I})'\sigma^2 + \mathbf{K}\boldsymbol{\Sigma}_{ee}\mathbf{K}'. \quad (8)$$

If model (1), (2), and (3) holds, and if $\boldsymbol{\Sigma}_{ee}$ and σ^2 are known, then replacing \mathbf{B} with

$$\tilde{\boldsymbol{\beta}} = \left(\mathbf{X}'\boldsymbol{\Sigma}_{zz}^{-1}\mathbf{X} \right)^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{zz}^{-1}\mathbf{Y} \quad (9)$$

and replacing \mathbf{G} with

$$\mathbf{G}_{zz} = \boldsymbol{\Sigma}_{zz}^{-1}\sigma^2 \quad (10)$$

in (4) defines the best linear unbiased predictor of \mathbf{y} . See Henderson (1950), Harville (1976), and Robinson (1991). If $\boldsymbol{\Sigma}_{ee}$ and σ^2 are also unknown, it is natural to use estimators of $\boldsymbol{\Sigma}_{ee}$ and σ^2 to construct an estimated best linear unbiased predictor. Very often, an estimator of $\boldsymbol{\Sigma}_{ee}$ is associated with the procedure used to construct the estimator \mathbf{Y} . Then σ^2 is estimated from model (1), (2), and (5), treating the estimator of $\boldsymbol{\Sigma}_{ee}$ as the true $\boldsymbol{\Sigma}_{ee}$.

One substitution predictor is

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\sigma}^2\hat{\boldsymbol{\Sigma}}_{zz}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}), \quad (11)$$

where

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\hat{\boldsymbol{\Sigma}}_{zz}^{-1}\mathbf{X} \right)^{-1}\mathbf{X}'\hat{\boldsymbol{\Sigma}}_{zz}^{-1}\mathbf{Y} \quad (12)$$

is the estimated generalized least squares estimator of $\boldsymbol{\beta}$,

$$\hat{\boldsymbol{\Sigma}}_{zz} = \mathbf{I}\hat{\sigma}^2 + \hat{\boldsymbol{\Sigma}}_{ee} \quad (13)$$

$\hat{\boldsymbol{\Sigma}}_{ee}$ is an estimator of $\boldsymbol{\Sigma}_{ee}$, and $\hat{\sigma}^2$ is an estimator of σ^2 . The estimator of σ^2 can be based on likelihood or analysis of variance procedures. Retaining only the terms in the Taylor expansion of the error in (11) that are errors in the basic estimators, we have

$$\begin{aligned} \hat{\mathbf{y}} - \mathbf{y} &\doteq \mathbf{e} - \mathbf{H}'\mathbf{z} + \mathbf{H}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &\quad + (\hat{\sigma}^2 - \sigma^2)\mathbf{H}'\boldsymbol{\Sigma}_{zz}^{-1}\mathbf{z} \\ &\quad - \mathbf{G}'(\hat{\boldsymbol{\Sigma}}_{ee} - \boldsymbol{\Sigma}_{ee})\boldsymbol{\Sigma}_{zz}^{-1}\mathbf{z}, \end{aligned} \quad (14)$$

where $\mathbf{H}' = \sum_{ee} \Sigma_{zz}^{-1}$ and $\mathbf{G}' = \mathbf{I} - \mathbf{H}' = \sigma^2 \Sigma_{zz}^{-1}$. If it is assumed that Σ_{ee} is distributed as a multiple of a Wishart matrix with d_e degrees of freedom, if the covariance between $\hat{\sigma}^2$ and $\hat{\Sigma}_{ee}$ is ignored, if expectations are computed as if $\hat{\sigma}^2$ and \mathbf{z} are independent, and if expectations are computed as if \mathbf{z} and $\hat{\Sigma}_{ee}$ are independent, an approximation to the variance of $\hat{\mathbf{y}} - \mathbf{y}$ obtained from (14) is

$$\mathbf{V}\{\hat{\mathbf{y}} - \mathbf{y}\} \doteq \Sigma_{ee} \mathbf{G} + \mathbf{H}' \mathbf{X} \mathbf{V}_{\beta\beta} \mathbf{X}' \mathbf{H} + \Gamma_{33} + \Gamma_{44}, \quad (15)$$

where

$$\mathbf{V}_{\beta\beta} = \mathbf{V}\{\hat{\beta}\} = (\mathbf{X}' \Sigma_{zz}^{-1} \mathbf{X})^{-1} + d_e^{-1} \text{tr}\{\Sigma_{zz}^{-1} \Sigma_{ee}\} \mathbf{L} \Sigma_{ee} \mathbf{L}',$$

$$\Gamma_{33} = \mathbf{H}' \Sigma_{zz}^{-1} \mathbf{H} V_{\sigma\sigma},$$

$$\Gamma_{44} = d_e^{-1} \sigma^4 \Sigma_{zz}^{-1} \Sigma_{ee} \Sigma_{zz}^{-1} \left[\text{tr}\{\Sigma_{zz}^{-1} \Sigma_{ee}\} \right],$$

$$\mathbf{L} = (\mathbf{X}' \Sigma_{zz}^{-1} \mathbf{X})^{-1} \mathbf{X}' \Sigma_{zz}^{-1}$$

and $V_{\sigma\sigma} = V\{\hat{\sigma}^2\}$ is the variance of $\hat{\sigma}^2$. The term $\Sigma_{ee} \mathbf{G}$ is the prediction covariance matrix if all parameters are known. The remaining three terms of (15) are the contributions to the variance due to estimating β , σ^2 , and Σ_{ee} , respectively. The second term in $\mathbf{V}\{\hat{\beta}\}$ is a crude approximation for the increase in the variance of $\hat{\beta}$ due to using an estimator of Σ_{zz} in place of Σ_{zz} in constructing $\hat{\beta}$.

If the dimension of Σ_{zz} is large and the degrees of freedom, d_e , only slightly larger than the dimension, then the second part of the variance of $\hat{\beta}$ and the term Γ_{44} can make important contributions to the variance. This is particularly true if σ^2 is small relative to the diagonal elements of Σ_{ee} . The Monte Carlo study of the next section demonstrates that the contribution to variance approximated by these terms can be important.

A predictor that reduces the effect of the estimation error in $\hat{\Sigma}_{ee}$ uses only diagonal elements of Σ_{ee} in the shrinkage component. Let

$$\hat{\mathbf{y}}_d = \mathbf{X} \hat{\beta}_d + \hat{\sigma}^2 \hat{\mathbf{D}}_{zz}^{-1} (\mathbf{Y} - \mathbf{X} \hat{\beta}_d), \quad (16)$$

where

$$\hat{\beta}_d = (\mathbf{X}' \hat{\mathbf{D}}_{zz}^{-1} \mathbf{X})^{-1} \mathbf{X}' \hat{\mathbf{D}}_{zz}^{-1} \mathbf{Y},$$

$$\hat{\mathbf{D}}_{zz} = \text{diag}(\hat{\Sigma}_{ee} + \mathbf{I} \hat{\sigma}^2),$$

$\hat{\sigma}^2$ is an estimator of σ^2 and $\text{diag}(\mathbf{A})$ is the diagonal matrix composed of the diagonal elements of \mathbf{A} . Retaining only the leading terms in the Taylor expansion of the error in (16) gives

$$\hat{\mathbf{y}}_d - \mathbf{y} \doteq -(\mathbf{w} - \mathbf{G}_d' \mathbf{z}) + \mathbf{H}_d' \mathbf{X} (\hat{\beta}_d - \beta)$$

$$+ (\hat{\sigma}^2 - \sigma^2) \mathbf{H}_d' \mathbf{D}_{zz}^{-1} \mathbf{z} - \mathbf{G}_d' (\hat{\mathbf{D}}_{ee} - \mathbf{D}_{ee}) \mathbf{D}_{zz}^{-1} \mathbf{z}, \quad (17)$$

where $\mathbf{D}_{zz} = \text{diag}\{\Sigma_{zz}\}$, $\mathbf{G}_d = \mathbf{D}_{zz}^{-1} \sigma^2$, $\mathbf{H}_d = \mathbf{I} - \mathbf{G}_d$, and $\mathbf{D}_{ee} = \text{diag}\{\Sigma_{ee}\}$. If \mathbf{w} and \mathbf{e} are normally distributed, and if $\hat{\sigma}^2$ and $\hat{\mathbf{D}}_{zz}$ are quadratic estimators, then $\hat{\sigma}^2$ and $\hat{\mathbf{D}}_{zz}$ are

uncorrelated with \mathbf{z} . The i -th element of $\mathbf{w} - \sigma^2 \mathbf{D}_{zz}^{-1} \mathbf{z}$ is uncorrelated with the i -th element of \mathbf{z} , but is not necessarily uncorrelated with the vector \mathbf{z} . If this possible correlation is ignored, if it is assumed that $\hat{\Sigma}_{ee}$ is a Wishart matrix with d_e degrees of freedom, and if the correlation between $\hat{\sigma}^2$ and $\hat{\Sigma}_{ee}$ is ignored, an approximation to the variance of $\hat{\mathbf{y}}_d - \mathbf{y}$ obtained from (17) is

$$\begin{aligned} \mathbf{V}\{\hat{\mathbf{y}}_d - \mathbf{y}\} &= \mathbf{H}_d' \mathbf{H}_d \sigma^2 + \mathbf{G}_d' \Sigma_{ee} \mathbf{G}_d + \mathbf{H}_d' \mathbf{X} \mathbf{V}_{\beta\beta} \mathbf{X}' \mathbf{H}_d \\ &\quad + \Gamma_{33dd} + \Gamma_{44dd}, \end{aligned} \quad (18)$$

where $\mathbf{G}_d = \mathbf{D}_{zz}^{-1} \sigma^2$, $\mathbf{H}_d = \mathbf{I} - \mathbf{G}_d$,

$$\mathbf{V}_{\beta\beta} = (\mathbf{X}' \mathbf{D}_{zz}^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{D}_{zz}^{-1} \Sigma_{zz} \mathbf{D}_{zz}^{-1} \mathbf{X} (\mathbf{X}' \mathbf{D}_{zz}^{-1} \mathbf{X})^{-1}, \quad (19)$$

$$\Gamma_{33dd} = \mathbf{H}_d' \mathbf{D}_{zz}^{-1} \Sigma_{zz} \mathbf{D}_{zz}^{-1} \mathbf{H}_d V_{\sigma\sigma}, \quad (20)$$

$$\Gamma_{44dd} = d_e^{-1} \mathbf{G}_d' \Omega \mathbf{G}_d$$

and the ij -th element of Ω is

$$\omega_{ij} = 2\sigma_{eeij} \sigma_{zzii}^{-1} \sigma_{zzjj}^{-1} \sigma_{zzij}.$$

The term in Γ_{44dd} is an estimator of the contribution to the variance due to using $\hat{\Sigma}_{ee}$ to estimate the covariance matrix. Expression (19) assumes that the contribution of the error in $\hat{\mathbf{D}}_{zz}$ to the variance of $\hat{\beta}$ can be ignored for large d_e . The difference between (15) and (18) is that the multipliers in (19) and (20) do not depend on the dimension of Σ_{zz} . Therefore, the error in estimating Σ_{zz} makes a smaller contribution to the variance. On the other hand, the variance of $\mathbf{w} - \mathbf{G}_d' \mathbf{z}$, the order one term of (17), will be larger than the corresponding term of the error in (14), unless Σ_{zz} is diagonal. The first two terms on the right of (18) are the variance of $\mathbf{w} - \mathbf{G}_d' \mathbf{z}$.

3. MONTE CARLO STUDY

To examine the variability in the predictors associated with variability in the estimation of Σ_{ee} we conducted a small Monte Carlo study. The model for the study is

$$\mathbf{Y}_j = \mu \mathbf{J} + \mathbf{w} + \mathbf{e}_j, \quad j = 1, 2, \dots, r \quad (21)$$

$$\mathbf{w} \sim (0, \mathbf{I} \sigma^2),$$

$$\mathbf{e}_j \sim \text{ind}(0, \Sigma_{ee}),$$

where \mathbf{J} is the k -dimensional column vector of ones, $\mathbf{J} = (1, 1, \dots, 1)'$, \mathbf{w} is the k -dimensional vector of random small area effects, \mathbf{e}_j is a vector of errors, and \mathbf{w} and \mathbf{e}_j are independent. The model is a simplified version of the model defined in (1), (2), and (3). The mean is the constant function and, hence, we use μ in place of β . To create a vector of correlated variables, we define, for $k = 8$,

$$\begin{bmatrix} e_{1j} \\ e_{2j} \\ e_{3j} \\ e_{4j} \\ e_{5j} \\ e_{6j} \\ e_{7j} \\ e_{8j} \end{bmatrix} = \begin{bmatrix} 1.3u_{1j} \\ 1.5u_{1j} + 0.4u_{2j} \\ 0.9u_{1j} + 0.9u_{3j} \\ 0.9u_{3j} + 1.6u_{4j} \\ 1.6u_{4j} + 0.6u_{5j} \\ 1.0u_{4j} + 1.6u_{6j} \\ 1.0u_{7j} \\ 2.83u_{8j} \end{bmatrix},$$

where u_{ij} are independent random variables. The w_i , $i = 1, 2, \dots, 8$, are $NI(0, 0.36)$ random variables, where $NI(\mu, \sigma^2)$ denotes normal independent random variables with mean μ and variance σ^2 . This configuration gives a range of error variances and a range of correlations between estimates.

The estimator of σ^2 used in the Monte Carlo study is

$$\hat{\sigma}^2 = \max \left\{ (k-1)^{-1} \times \left[(\bar{\mathbf{y}} - \mathbf{J}\hat{\mu}_{(0)})' (\bar{\mathbf{y}} - \mathbf{J}\hat{\mu}_{(0)}) - \text{tr} \left\{ r^{-1} \hat{\Sigma}_{ee} \mathbf{A}_0 \right\} \right], 0 \right\} \quad (22)$$

where $\text{tr}[\mathbf{A}]$ is the trace of the matrix \mathbf{A} ,

$$\mathbf{A}_0 = \mathbf{I} - k^{-1} \mathbf{J}\mathbf{J}'$$

$$\hat{\Sigma}_{ee} = (r-1)^{-1} \sum_{j=1}^r (\mathbf{Y}_j - \bar{\mathbf{y}})(\mathbf{Y}_j - \bar{\mathbf{y}})', \quad (23)$$

and

$$\hat{\mu}_0 = k^{-1} \mathbf{J}' \bar{\mathbf{y}}. \quad (24)$$

The estimator $\hat{\sigma}^2$ is a quadratic estimator closely related to the analysis of variance estimator.

Two predictors were compared in the Monte Carlo study. Both are of the form

$$\hat{\mathbf{y}} = \bar{\mathbf{y}} - \hat{\mathbf{H}}' (\bar{\mathbf{y}} - \hat{\mu} \mathbf{J}), \quad (25)$$

where

$$\bar{\mathbf{y}} = r^{-1} \sum_{j=1}^r \mathbf{Y}_j.$$

They differ in the construction of $\hat{\mathbf{H}}$ and $\hat{\mu}$. The first predictor is of the form (11) and uses the full estimated $\hat{\Sigma}_{ee}$ in $\hat{\mathbf{H}}$ and in the estimator of μ . The predictor is called the general predictor as an abbreviation for estimated generalized least squares predictor. The general predictor is

$$\hat{\mathbf{y}}_g = \bar{\mathbf{y}} - \hat{\mathbf{H}}'_g (\bar{\mathbf{y}} - \hat{\mu}_g \mathbf{J}), \quad (26)$$

where

$$\hat{\mathbf{H}}'_g = r^{-1} \hat{\Sigma}_{ee} \hat{\Sigma}_{zz}^{-1},$$

$$\hat{\mu}_g = (\mathbf{J}' \hat{\Sigma}_{zz}^{-1} \mathbf{J})^{-1} \mathbf{J}' \hat{\Sigma}_{zz}^{-1} \bar{\mathbf{y}}, \quad (27)$$

$$\hat{\Sigma}_{zz} = r^{-1} \hat{\Sigma}_{ee} + \mathbf{I} \hat{\sigma}^2, \quad (28)$$

and $\hat{\mu}_g$ is the estimated generalized least squares estimator of μ .

The second predictor is

$$\hat{\mathbf{y}}_d = \bar{\mathbf{y}} - \hat{\mathbf{H}}'_d (\bar{\mathbf{y}} - \hat{\mu}_d \mathbf{J}), \quad (29)$$

where

$$\hat{\mathbf{H}}'_d = r^{-1} \mathbf{M}_{ee} \hat{\Sigma}_{zz}^{-1},$$

$\mathbf{M}_{ee} = \text{diag} \hat{\Sigma}_{ee}$, $\hat{\Sigma}_{zz} = \text{diag} \hat{\Sigma}_{zz}$, and the estimated μ is

$$\hat{\mu}_d = [\mathbf{J}' \hat{\Sigma}_{zz}^{-1} \mathbf{J}]^{-1} \mathbf{J}' \hat{\Sigma}_{zz}^{-1} \bar{\mathbf{y}}.$$

This predictor might be called the diagonal predictor because only the diagonal elements of $\hat{\Sigma}_{ee}$ are used in the construction.

The entries in Table 1 are for $r = 14$. Each sample is composed of a random selection of \mathbf{w} and a random sample of 14 \mathbf{e} -vectors. Results are given for errors $u_{ij} \sim NI(0, 2)$ and errors that are centered one-degree-of-freedom chi-square random variables. Thus, in both cases the errors have zero means and variances equal to two. The mean μ was set equal to zero. The second column of Table 1 contains the variance of the sample mean as an estimator of the w_i . Column three of Table 1 contains the ratio of the Monte Carlo variance of an element of $\hat{\mathbf{y}}_g$, where $\hat{\mathbf{y}}_g$ is defined by (28), to the Monte Carlo variance of the corresponding element of $\bar{\mathbf{y}}$ for normal errors. The ratios for elements one through four and element 7 are greater than one. The last two elements of \mathbf{Y}_j are uncorrelated with other elements. Element seven has a small variance and element eight has a large variance. There is a large loss for the predictor relative to the simple mean for element seven and a large gain for element eight.

The fourth column of Table 1 contains the ratios of the variance of the predictor of (29) to the variance of the mean for normal errors. In all cases the diagonal predictor is superior to the general predictor defined in (28). The difference is relatively constant at about 30%. The diagonal predictor is not always superior to the simple mean but the loss is small for elements one, three, and seven. On the other hand, the gains relative to the simple mean are large for elements six and eight. The Monte Carlo variances for both predictors are larger than the approximations associated with equations (15) and (18) except for element 8.

It is somewhat surprising that the diagonal procedure did better relative to the simple mean for chi-square errors than for normal errors. With the chi-square error, the estimated mean and estimated variance are correlated. Hence, on the average, the large positive mean deviations are pulled toward the mean by a larger amount than the smaller negative deviation. The Associate Editor conjectured, and we concur, that this is one reason for the superior performance of the diagonal predictor. On the other hand, the general

prediction procedure is poorer relative to the simple mean for chi-square errors than for normal errors. As the last column of Table 1 demonstrates, the diagonal predictor procedure uniformly dominates both the mean and the general prediction procedure for this parametric configuration with chi-square errors.

Table 1
Monte Carlo Variance Ratios for Alternative
Small Area Predictors
(10,000 samples, $r = 14$)

i	$V\{\bar{y}_i - w_i\}$	Normal Errors		Chi-square Errors	
		$\hat{V}\{\hat{y}_{gi} - w_i\}$ $\hat{V}\{\bar{y}_i - w_i\}$	$\hat{V}\{\hat{y}_{di} - w_i\}$ $\hat{V}\{\bar{y}_i - w_i\}$	$\hat{V}\{\hat{y}_{gi} - w_i\}$ $\hat{V}\{\bar{y}_i - w_i\}$	$\hat{V}\{\hat{y}_{di} - w_i\}$ $\hat{V}\{\bar{y}_i - w_i\}$
1	0.2414	1.277	1.025	1.430	0.899
2	0.3445	1.252	0.875	1.371	0.768
3	0.2268	1.351	1.019	1.480	0.954
4	0.4771	1.003	0.735	1.099	0.686
5	0.4113	0.926	0.876	1.016	0.699
6	0.5121	0.913	0.677	0.975	0.618
7	0.1449	1.366	1.006	2.261	0.896
8	1.1214	0.520	0.384	0.725	0.371

The Monte Carlo variances of $\hat{\mu}_0$, $\hat{\mu}_g$, and $\hat{\mu}_d$ as estimators of μ are 0.150, 0.273, and 0.146, respectively. If Σ_{ee} and σ^2 are known, the variances of $\hat{\mu}_0$, $\hat{\mu}_g$, and $\hat{\mu}_d$ are 0.149, 0.122, and 0.140, respectively. The use of an estimated covariance matrix for $\hat{\mu}_g$ produced an estimator with larger variance than that of the simple mean.

The predictors are unbiased under the model when the errors are normally distributed. The predictors are biased with chi-square errors because the sample mean is correlated with the sample variance. Table 2 contains the Monte Carlo bias divided by the Monte Carlo standard error of the mean. The bias of the general procedure is 20% to 50% larger than that of the diagonal procedure. In both cases, the squared bias added to the variance produces a mean square error for the procedure that is about 4% to 10% larger than the variance.

This small study demonstrates that use of an estimated covariance matrix with large variability can lead to predictors that are less efficient than the simple mean.

Table 2
Monte Carlo Relative Bias of Alternative
Small Area Predictors
(10,000 samples, $r = 14$, chi-square errors)

i	Ave. $(\hat{y}_{gi} - w_i)$ $[\hat{V}\{\bar{y}_i - w_i\}]^{1/2}$	Ave. $(\hat{y}_{di} - w_i)$ $[\hat{V}\{\bar{y}_i - w_i\}]^{1/2}$
1	-0.28	-0.19
2	-0.27	-0.18
3	-0.30	-0.17
4	-0.27	-0.18
5	-0.26	-0.21
6	-0.29	-0.20
7	-0.24	-0.20
8	-0.24	-0.21

4. APPLICATION TO PES DATA FOR POSTCENSAL ESTIMATION

4.1 Postcensal Estimation

The U.S. Bureau of the Census provides annual estimates of the population of small areas based on the decennial censuses and on other sources of information. To consider the possible use of adjusted 1990 Census counts in the postcensal estimation process, the Bureau examined the PES data and defined a new set of 357 poststrata.

The 357 poststrata are composed of 51 poststratum groups, each of which is subdivided into 7 age-sex categories. The seven age-sex categories were (1) both sexes 0-17, (2) males 18-29, (3) males 30-49, (4) females 18-29, (5) females 30-49, (6) males 50+ and (7) females 50+. The factors that define the 51 poststratum groups are race/ethnicity (Non-Hispanic White, Black, Non-Black Hispanic, Asian, American Indian); tenure (owner, renter); type of area (urbanized area of population greater than 250,000, other urbanized area, non-urbanized area) and region (West, South, Midwest, Northeast). Due to sample size limitations, American Indians comprised a single poststratum group and Asians were dichotomized into two poststratum groups – owners and renters. Of the remaining 48 poststratum groups, the first 24 groups reflect a full cross classification of categories for Non-Hispanic White. The next 12 groups are for Black and provide a full cross classification of tenure by region for urbanized areas of population greater than 250,000 but otherwise do not provide regional detail. The same 12 poststratum groups were used for Non-Black Hispanics as were used for Blacks.

A 357×357 covariance matrix was obtained with the same jackknife algorithm used for the 1392 poststrata of the 1990 PES. We denote this raw covariance matrix by Σ_{ee} . Hogan (1993) provides a detailed description of the 357 poststrata and gives the motivation for their construction.

4.2 Regression Model

We eliminated Asian and American Indian data from the smoothing process. Hence, minority refers to the combination of Black and Non-Black Hispanic. The data set of interest contains 336 adjustment factors and their estimated raw covariances. The minority by age-sex interaction was included in the regression model after examination of the 1990 data indicated that the net undercount differential between Black and Non-Black varied by sex and age-group. The regression model (1) contains 21 explanatory variables. They are:

1. X_0 = intercept
2. X_j = indicator variable for age-sex categories:
 $j = 1, 2, \dots, 6$ in the order; ages 0-17, male 18-29, male 30-49, etc. (female 50+ is the class with no variable)

3. X_7 = indicator variable for renter
4. X_8 = indicator variable for Black
5. X_9 = indicator variable for Non-Black Hispanic
6. X_j = indicator variable for type of place: $j = 10, 11$ for urbanized area 250,000+ and other urban, respectively
7. X_j = indicator variable for region: $j = 12, 13, 14$ for Northeast, South and West, respectively
8. X_j = indicator variable for minority by age-sex interaction: $j = 15, \dots, 20$ for minority 0-17, minority male (18-29), etc.

The variables X_{12} , X_{13} and X_{14} were the 1990 census proportions of persons in the poststratum group in the particular region for the Black and Non-Black Hispanic poststratum groups that were combined over regions.

A refinement was made in model (3) for the empirical application. On the basis of preliminary analysis, the specified error structure of \mathbf{w} , the model error, was changed from $\Sigma_{ww} = \sigma^2 \mathbf{I}$ to

$$\Sigma_{ww} = \mathbf{K}_1 \sigma_1^2 + \mathbf{K}_2 \sigma_2^2, \quad (30)$$

where \mathbf{K}_1 is an $n \times n$ diagonal matrix with ones for minority poststrata and zeros elsewhere and \mathbf{K}_2 is an $n \times n$ diagonal matrix with ones for nonminority poststrata and zeros elsewhere. The estimated variances are $\hat{\sigma}_1^2 = 0.000506$ (0.000140) and $\hat{\sigma}_2^2 = 0.000112$ (0.000030), where the numbers in parentheses are standard errors. The standard error of the difference is (0.000141). Hence there is evidence that the variances are different for the two groups.

In our discussion of predictors, we considered two predictors, the substitution predictor of (11) and the diagonal predictor of (16). It is natural to consider a compromise predictor of the form

$$\begin{aligned} \hat{\mathbf{y}}_\varphi &= \mathbf{X} \hat{\boldsymbol{\beta}}_\varphi + \hat{\mathbf{G}}_\varphi' (\mathbf{Y} - \mathbf{X} \hat{\boldsymbol{\beta}}_\varphi) \\ &= \mathbf{Y} - \hat{\mathbf{H}}_\varphi' (\mathbf{Y} - \mathbf{X} \hat{\boldsymbol{\beta}}_\varphi), \end{aligned} \quad (31)$$

where $0 \leq \varphi \leq 1$,

$$\hat{\mathbf{G}}_\varphi = \hat{\Sigma}_{\varphi\varphi}^{-1} \hat{\Sigma}_{\varphi ww},$$

$$\hat{\mathbf{H}}_\varphi = \mathbf{I} - \hat{\mathbf{G}}_\varphi = \hat{\Sigma}_{\varphi\varphi}^{-1} [\varphi \hat{\mathbf{D}}_{ee} + (1 - \varphi) \hat{\Sigma}_{ee}],$$

$$\hat{\Sigma}_{\varphi\varphi} = \hat{\Sigma}_{ww} + \varphi \hat{\mathbf{D}}_{ee} + (1 - \varphi) \hat{\Sigma}_{ee},$$

$$\hat{\mathbf{D}}_{ee} = \text{diag} \{ \hat{\Sigma}_{ee} \},$$

$$\hat{\boldsymbol{\beta}}_\varphi = (\mathbf{X}' \hat{\Sigma}_{\varphi\varphi}^{-1} \mathbf{X})^{-1} \mathbf{X}' \hat{\Sigma}_{\varphi\varphi}^{-1} \mathbf{Y},$$

and

$$\hat{\Sigma}_{ww} = \mathbf{K}_1 \hat{\sigma}_1^2 + \mathbf{K}_2 \hat{\sigma}_2^2.$$

The predictor (31) with $\varphi = 0$ is the substitution predictor and the predictor (31) with $\varphi = 1$ the diagonal predictor. There should be some φ , $0 < \varphi < 1$, that gives a predictor with smaller prediction variance than either of the extremes.

The PES direct estimate of the total number of persons is the weighted sum of the adjustment factors, where the weights are the census counts in the post strata. The standard error of the direct estimator of the total is relatively small and the direct estimator is judged to be the preferred estimator of the total. Therefore, the model predictors are constructed subject to the constraint that the weighted sum of the predictors is equal to the direct estimate of the total. Thus, the restriction is

$$\hat{Y}_T = \sum_{i=1}^{336} a_i y_i = \sum_{i=1}^{336} a_i \tilde{y}_i,$$

where \hat{Y}_T is PES direct estimator of the total, a_i is the census count in the i -th post stratum, and \tilde{y}_i is the final predictor. In the actual computations the a_i were normalized to sum to one. Battese, Harter and Fuller (1988) made an adjustment in the predictions to create estimators to meet the restriction. Ghosh and Rao (1994) discuss such adjustments. We use a procedure that permits direct estimation of the variance of the restricted predictions.

We imposed the restriction on the initial predictors by a procedure that, approximately, constructed the best predictors of 335 quantities that are estimated to be uncorrelated with \hat{Y}_T . Let $\hat{\Sigma}_{\varphi\varphi}$ be the estimated covariance matrix of $\mathbf{Y} = (Y_1, Y_2, \dots, Y_{336})'$ and define

$$\mathbf{CY} = (\hat{Y}_T, Y_2 - b_2 \hat{Y}_T, \dots, Y_{336} - b_{336} \hat{Y}_T)',$$

where

$$\mathbf{C} = \mathbf{BT},$$

$$\mathbf{T} = \begin{pmatrix} a & \\ \mathbf{0} & \mathbf{I}_{335} \end{pmatrix},$$

$$\mathbf{a} = (a_1, a_2, \dots, a_{336}),$$

$$\mathbf{B} = \begin{pmatrix} 1 & \mathbf{0}' \\ -b_{335} & \mathbf{I}_{335} \end{pmatrix},$$

$$\mathbf{b}_{335} = \begin{pmatrix} \mathbf{0}' \\ \mathbf{I}_{335} \end{pmatrix}' \hat{\Sigma}_{\varphi\varphi}^{-1} \mathbf{a}' (\mathbf{a}' \hat{\Sigma}_{\varphi\varphi}^{-1} \mathbf{a})^{-1},$$

\mathbf{I}_k is the $k \times k$ identity matrix, and $\mathbf{0}$ is a column vector containing all zeros. The elements of \mathbf{CY} are uncorrelated with \hat{Y}_T .

If we let $\hat{\mathbf{y}}$ be the model predictor of \mathbf{y} , then the model predictor of \mathbf{Cy} is $\mathbf{C}\hat{\mathbf{y}}$. If we use the model predictor for the last 335 elements of \mathbf{Cy} and use \hat{Y}_T as the estimator for the first element of \mathbf{Cy} , the predictor of \mathbf{y} is

$$\tilde{\mathbf{y}} = \mathbf{Y} - \mathbf{C}^{-1}\mathbf{A}\hat{\mathbf{C}}\hat{\mathbf{H}}_{\phi}'(\mathbf{Y} - \mathbf{X}\hat{\beta}_{\phi}), \quad (32)$$

where

$$\mathbf{A} = \begin{pmatrix} 0 & \mathbf{0}' \\ \mathbf{0} & \mathbf{I}_{335} \end{pmatrix}.$$

The estimated variance of $\tilde{\mathbf{y}}$ is

$$\begin{aligned} & \hat{\mathbf{V}}\{\tilde{\mathbf{y}} - \mathbf{y}\} \\ &= (\mathbf{I} - \hat{\mathbf{H}}_{\phi}')\hat{\Sigma}_{ee}(\mathbf{I} - \hat{\mathbf{H}}_{\phi}')' + \hat{\mathbf{H}}_{\phi}'\hat{\Sigma}_{ww}\hat{\mathbf{H}}_{\phi} \\ &+ \mathbf{C}^{-1}\mathbf{A}\hat{\mathbf{C}}[\hat{\mathbf{H}}_{\phi}'\mathbf{X}'\hat{\mathbf{V}}_{\beta\beta}\mathbf{X}'\hat{\mathbf{H}}_{\phi} + \hat{\Gamma}_{33} + \hat{\Gamma}_{44}]\mathbf{C}'\mathbf{A}\mathbf{C}^{-1}', \quad (33) \end{aligned}$$

where $\hat{\mathbf{H}}_{\phi}' = \mathbf{C}^{-1}\mathbf{A}\hat{\mathbf{C}}\hat{\mathbf{H}}_{\phi}'$, and $\hat{\mathbf{V}}_{\beta\beta}$, $\hat{\Gamma}_{33}$ and $\hat{\Gamma}_{44}$ are defined in Appendix B. The sum of the first two terms on the right of (33) is an estimator of the variance treating $\hat{\mathbf{H}}_{\phi}$ as a fixed matrix. The final term on the right of (33) estimates the increase in variance due to estimating the variance.

4.3 Smoothed Factors

For the vector of 336 observations, we produced smoothed factors using the generalized predictor (32) for several values of ϕ . Note that $\phi = 0$ corresponds to the substitution predictor and $\phi = 1$ corresponds to the diagonal predictor.

The estimated standard errors of the predictors were calculated using the crude variance approximation of Appendix B. The average of the ratios of the standard error of \tilde{y}_{ϕ} to $\tilde{y}_{0.6}$ for some selected values of ϕ are given in Table 3. The ordering of the ratios is approximately the same for the 48 stratum groups as for the original 336 poststrata. A poststratum group is formed by combining the seven age-sex cells within a given race-by-tenure-by-urbanity-by-region classification. On the basis of these calculations, a ϕ of 0.5 or 0.6 is the preferred estimator, although the estimated differences in efficiencies are not large. Any member of the ϕ -class is much superior to the original Y -estimator. The average estimated variance efficiency is about 400% for the ϕ -predictors, relative to the original poststratum estimators.

Table 3
Average of Ratio of Standard Error of \tilde{y}_{ϕ}
and of Y to Standard Error of $\tilde{y}_{0.6}$

Predictor	336	48
	Poststrata	Poststratum groups
$\phi = 0$	1.014	1.045
$\phi = 0.5$	0.995	1.001
$\phi = 0.6$	1.000	1.000
$\phi = 0.7$	1.006	1.001
$\phi = 0.8$	1.014	1.005
$\phi = 1.0$	1.046	1.037
Original Y	2.235	2.294

Table 4 presents the raw PES estimates, \mathbf{Y} , and the $\tilde{y}_{0.6}$ estimates of net undercount for each of 48 poststratum groups. The net undercount is the difference between the estimated total population in the poststratum and the census count divided by the census count.

We chose $\phi = 0.6$ as the preferred estimator on the basis of the crude standard error ratios of Table 3. The predictions and standard errors are very similar for $\phi = 0.5$, 0.6 and 0.7. A ϕ greater than zero has advantages over a ϕ of zero. The accuracy of the numerical calculations should be better with ϕ greater than zero because $\hat{\Sigma}_{\phi\phi}$ has larger eigenvalues with $\phi > 0$ than with $\phi = 0$. One could make a case for using $\phi = 1.0$ because of the simplicity of the calculations and of the good estimated relative efficiency.

The estimated standard errors of the predictors are considerably smaller than those of the raw estimates. In addition, the set of predictors contains fewer extreme estimates. For example, for poststratum groups 34, 39 and 48, the $\tilde{y}_{0.6}$ estimates of the percent net undercount are 6.04, 0.17 and 7.51 while the raw estimates are 11.06, -4.14 and 18.76, respectively. Most smoothed estimates differ from the direct estimate by less than one direct estimated standard error. The three largest standardized differences are for Black Owner-Large Urban in the West, Black Renter-Large Urban in the Northeast, and Non-Black Hispanic Owner-Large Urban in the Midwest. In the three cases, the difference between the direct estimate and the smoothed estimate divided by the direct standard error is about 1.8.

ACKNOWLEDGEMENTS

This research was partly supported by Cooperative Agreement 43-3AEU-3-80088 between Iowa State University, the National Agricultural Statistics Service and the U.S. Bureau of the Census. This paper reports the results of research and analysis undertaken by staff of the Bureau of the Census and Iowa State University. It has undergone a more limited review than official Census Bureau publications. This paper is released to inform interested parties of research and to encourage discussion. We thank the referees and editors for many comments that led to improvements in the manuscript.

Table 4
Estimated Percent Net Undercount by Poststratum Group

Poststratum Group	\bar{Y}	s.e. (\bar{Y})	$\bar{Y}_{0.6}$	s.e. ($\bar{Y}_{0.6}$)
Non-Hispanic White Owner Large Urban				
1. N.E.	-2.08	1.04	-0.63	0.60
2. South	0.69	0.72	0.38	0.44
3. Midwest	-0.26	0.39	-0.13	0.31
4. West	-0.34	0.64	-0.02	0.44
Non-Hispanic White Owner Other Urban				
5. N.E.	-1.07	0.48	-0.73	0.35
6. South	0.52	0.43	0.53	0.33
7. Midwest	-0.10	0.40	0.01	0.31
8. West	0.63	0.58	0.30	0.40
Non-Hispanic White Owner Non-Urban				
9. N.E.	-0.53	0.69	-0.28	0.47
10. South	0.18	0.69	0.58	0.45
11. Midwest	-0.70	1.16	0.16	0.64
12. West	0.29	0.69	0.38	0.46
Non-Hispanic White Renter Large Urban				
13. N.E.	1.17	1.43	2.07	0.61
14. South	2.62	1.56	3.53	0.64
15. Midwest	2.39	1.70	2.53	0.60
16. West	3.28	1.72	3.10	0.58
Non-Hispanic White Renter Other Urban				
17. N.E.	3.53	1.62	2.29	0.61
18. South	3.30	1.86	3.67	0.67
19. Midwest	1.24	1.13	2.39	0.53
20. West	4.70	1.47	3.20	0.57
Non-Hispanic White Renter Non- Urban				
21. N.E.	6.97	4.67	3.54	0.92
22. South	6.65	1.93	3.60	0.66
23. Midwest	2.93	1.60	2.36	0.66
24. West	6.48	2.06	3.48	0.67
Black Owner Large Urban				
25. N.E.	1.65	1.96	0.97	0.91
26. South	2.20	0.94	2.30	0.70
27. Midwest	0.82	0.88	1.13	0.67
28. West	6.49	2.16	2.54	0.96
Black Owner Other Urban				
29. U.S.	1.36	1.01	2.05	0.72
Black Owner Non- Urban				
30. U.S.	3.64	2.03	2.85	0.98
Black Renter Large Urban				
31. N.E.	9.13	1.93	5.57	0.96
32. South	6.69	2.17	6.42	1.10
33. Midwest	6.38	1.91	5.43	1.03
34. West	11.06	3.35	6.04	1.12
Black Renter Other Urban				
35. U.S.	4.33	1.28	4.99	0.82
Black Renter Non- Urban				
36. U.S.	4.84	5.95	5.90	1.24
Non-Black Hispanic Owner Large Urban				
37. N.E.	0.68	4.44	3.00	1.18
38. South	2.59	0.95	2.52	0.72
39. Midwest	-4.14	2.38	0.17	0.97
40. West	2.98	0.92	2.89	0.68
Non-Black Hispanic Owner Other Urban				
41. U.S.	0.95	1.70	2.32	0.87
Non-Black Hispanic Owner Non-Urban				
42. U.S.	2.80	2.83	2.88	1.16
Non-Black Hispanic Renter Large Urban				
43. N.E.	7.21	4.04	5.85	1.27
44. South	10.30	3.11	7.35	1.15
45. Midwest	7.11	3.74	5.71	1.21
46. West	6.29	2.09	6.45	0.98
Non-Black Hispanic Renter Other Urban				
47. U.S.	7.07	3.10	6.26	1.09
Non-Black Hispanic Renter Non- Urban				
48. U.S.	18.76	7.24	7.51	1.38

APPENDIX A: Estimation of Σ_{ww}

The estimators of σ_1^2 and σ_2^2 of Σ_{ww} are patterned after analysis of variance estimators. The estimation process contains several steps using improved estimators from one step in the next step. We partition the regression problem as

$$\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{pmatrix},$$

where $(\mathbf{Y}_1, \mathbf{X}_1)$ contains the observations for minorities and $(\mathbf{Y}_2, \mathbf{X}_2)$ contains the remaining observations. Let \mathbf{Y}_1 be an n_1 -dimensional column vector and \mathbf{Y}_2 be an n_2 -dimensional column vector observations. An initial estimator of $(\beta_1', \beta_2')'$ is

$$\begin{pmatrix} \tilde{\beta}_1 \\ \tilde{\beta}_2 \end{pmatrix} = \begin{pmatrix} (\mathbf{X}_1' \hat{\Sigma}_{ee11}^{-1} \mathbf{X}_1)^{-1} & \mathbf{X}_1' \hat{\Sigma}_{ee11}^{-1} \mathbf{Y}_1 \\ (\mathbf{X}_2' \hat{\Sigma}_{ee22}^{-1} \mathbf{X}_2)^{-1} & \mathbf{X}_2' \hat{\Sigma}_{ee22}^{-1} \mathbf{Y}_2 \end{pmatrix},$$

where

$$\hat{\Sigma}_{ee} = \begin{pmatrix} \hat{\Sigma}_{ee11} & \hat{\Sigma}_{ee12} \\ \hat{\Sigma}_{ee21} & \hat{\Sigma}_{ee22} \end{pmatrix}$$

is partitioned to conform to the partition of \mathbf{Y} .

Initial estimators of σ_1^2 and σ_2^2 are

$$\hat{\sigma}_i^2 = \max \left\{ \left[(\mathbf{Y}_i - \mathbf{X}_i \tilde{\beta}_i)' \hat{\Sigma}_{eeii}^{-1} (\mathbf{Y}_i - \mathbf{X}_i \tilde{\beta}_i) - g_{2i} \right] g_{2i}^{-1}, 0 \right\},$$

for $i = 1, 2$, where

$$g_{1i} = \text{tr} \left\{ \hat{\Sigma}_{ee11} (\mathbf{I}_{n_i} - \mathbf{X}_i \mathbf{A}_{Mii})' \hat{\Sigma}_{eeii}^{-1} (\mathbf{I}_{n_i} - \mathbf{X}_i \mathbf{A}_{Mii}) \right\},$$

$$g_{2i} = \text{tr} \left\{ (\mathbf{I}_{n_i} - \mathbf{X}_i \mathbf{A}_{Mii})' \hat{\Sigma}_{eeii}^{-1} (\mathbf{I}_{n_i} - \mathbf{X}_i \mathbf{A}_{Mii}) \right\},$$

$$\mathbf{A}_{Mii} = (\mathbf{X}_i' \hat{\Sigma}_{eeii}^{-1} \mathbf{X}_i)^{-1} \mathbf{X}_i' \hat{\Sigma}_{eeii}^{-1},$$

and \mathbf{I}_{n_i} is the $n_i \times n_i$ identity matrix.

The final estimators are

$$\hat{\sigma}_i^2 = \max \left\{ \left[(\mathbf{Y}_i - \mathbf{X}_i \tilde{\beta}_i)' \hat{\Sigma}_{zzii}^{-1} (\mathbf{Y}_i - \mathbf{X}_i \tilde{\beta}_i) - \tilde{g}_{2i} \right] \tilde{g}_{2i}^{-1}, 0 \right\},$$

for $i = 1, 2$, where

$$\tilde{g}_{1i} = \text{tr} \left\{ \hat{\Sigma}_{eeii} (\mathbf{I}_{n_i} - \mathbf{X}_i \mathbf{A}_{Mii})' \hat{\Sigma}_{zzii}^{-1} (\mathbf{I}_{n_i} - \mathbf{X}_i \mathbf{A}_{Mii}) \right\}$$

$$\tilde{g}_{2i} = \text{tr} \left\{ (\mathbf{I}_{n_i} - \mathbf{X}_i \mathbf{A}_{Mii})' \hat{\Sigma}_{zzii}^{-1} (\mathbf{I}_{n_i} - \mathbf{X}_i \mathbf{A}_{Mii}) \right\}$$

$$\hat{\Sigma}_{zzii} = \hat{\Sigma}_{eeii} + \hat{\sigma}_i^2 \mathbf{I}_{n_i}$$

$$\tilde{\beta}_i = (\mathbf{X}_i' \hat{\Sigma}_{zzii}^{-1} \mathbf{X}_i)^{-1} \mathbf{X}_i' \hat{\Sigma}_{zzii}^{-1} \mathbf{Y}_i = \mathbf{A}_{Mii}^{-1} \mathbf{Y}_i,$$

Estimators of the variance are

$$\begin{aligned} \hat{V}\{\hat{\sigma}_i^2\} &= 2\hat{g}_{2i}^{-2} \\ &\times \text{tr} \left\{ \left[\hat{\Sigma}_{zzii} (\mathbf{I}_{n_i} - \mathbf{X}_i \mathbf{A}_{Mii})' \hat{\Sigma}_{zzii}^{-1} (\mathbf{I}_{n_i} - \mathbf{X}_i \mathbf{A}_{Mii}) \right]^2 \right\} \\ &+ 2\hat{g}_{2i}^{-2} d_e^{-1} \\ &\times \text{tr} \left\{ \left[\hat{\Sigma}_{eeii} (\mathbf{I}_{n_i} - \mathbf{X}_i \mathbf{A}_{Mii})' \hat{\Sigma}_{zzii}^{-1} (\mathbf{I}_{n_i} - \mathbf{X}_i \mathbf{A}_{Mii}) \right]^2 \right\}, \end{aligned}$$

for $i = 1, 2$. The estimated covariance is

$$\begin{aligned} \hat{C}\{\hat{\sigma}_1^2, \hat{\sigma}_2^2\} &= 2\text{tr} \left\{ \hat{\Sigma}_{zz} \mathbf{M}_{11} \hat{\Sigma}_{zz} \mathbf{M}_{22} \right\} \\ &+ 2d_e^{-1} \text{tr} \left\{ \hat{\Sigma}_{ee} \mathbf{M}_{11} \hat{\Sigma}_{ee} \mathbf{M}_{22} \right\}, \end{aligned}$$

where

$$\mathbf{M}_{11} = \begin{pmatrix} g_{21}^{-1} (\mathbf{I}_{n_1} - \mathbf{X}_1 \mathbf{A}_{M11})' \hat{\Sigma}_{zz11}^{-1} (\mathbf{I}_{n_1} - \mathbf{X}_1 \mathbf{A}_{M11}) & \mathbf{0} \\ \mathbf{0}' & \mathbf{0} \end{pmatrix}$$

and

$$\mathbf{M}_{22} = \begin{pmatrix} \mathbf{0} & \mathbf{0}' \\ \mathbf{0} & g_{22}^{-1} (\mathbf{I}_{n_2} - \mathbf{X}_2 \mathbf{A}_{M22})' \hat{\Sigma}_{zz22}^{-1} (\mathbf{I}_{n_2} - \mathbf{X}_2 \mathbf{A}_{M22}) \end{pmatrix}.$$

See Searle (1971, Chapter 2 and p. 435).

APPENDIX B: Approximations for the Variance of Predictors

Our model is

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{w} + \mathbf{e}, \quad (\text{B.1})$$

where \mathbf{Y} is an n -dimensional column vector, \mathbf{X} is an $n \times k$ fixed matrix,

$$\begin{pmatrix} \mathbf{w} \\ \mathbf{e} \end{pmatrix} \sim N \left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \Sigma_{ww} & \mathbf{0} \\ \mathbf{0} & \Sigma_{ee} \end{pmatrix} \right), \quad (\text{B.2})$$

and Σ_{ww} is defined in (30) of the text.

For purpose of variance estimation, we assume $\hat{\Sigma}_{ee}$ is an unbiased estimator of Σ_{ee} distributed as a multiple of a Wishart matrix with d_e degrees of freedom independent of (\mathbf{w}, \mathbf{e}) . We let \mathbf{y} be the unknown true vector to be predicted and write

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{w} \quad \text{and} \quad \mathbf{z} = \mathbf{w} + \mathbf{e}.$$

By a Taylor expansion

$$\begin{aligned} \hat{\mathbf{y}}_\varphi - \mathbf{y} &= \mathbf{e} - \hat{\mathbf{H}}_\varphi' (\mathbf{Y} - \mathbf{X}\hat{\beta}_\varphi) \\ &= \mathbf{e} - \mathbf{H}_\varphi' \mathbf{z} + \mathbf{H}_\varphi' \mathbf{X} (\hat{\beta}_\varphi' - \beta) - (\hat{\mathbf{H}}_\varphi' - \mathbf{H}_\varphi') \mathbf{z} + O_p(n^{-1}), \quad (\text{B.3}) \end{aligned}$$

where

$$\mathbf{H}_\varphi = \Sigma_{\varphi\varphi}^{-1} [\varphi \mathbf{D}_{ee} + (1 - \varphi) \Sigma_{ee}]$$

and $\hat{\mathbf{H}} = \hat{\mathbf{H}}_\varphi$ is defined in (31). The error in $\hat{\beta}_\varphi$ is

$$\hat{\beta}_\varphi - \beta = (\mathbf{X}' \hat{\Sigma}_{\varphi\varphi}^{-1} \mathbf{X})^{-1} \mathbf{X}' \hat{\Sigma}_{\varphi\varphi}^{-1} \mathbf{z} \quad (\text{B.4})$$

Now $\hat{\Sigma}_{ee}$ is independent of \mathbf{z} and $\mathbf{Y} - \mathbf{X}\tilde{\beta}$ is uncorrelated with $\tilde{\beta} - \beta$ if the true Σ_{zz} is used in place of $\hat{\Sigma}_{\varphi\varphi}$. Therefore

$$E\{\mathbf{H}_0' \mathbf{X}(\tilde{\beta}_0 - \beta) \mathbf{z}' (\hat{\mathbf{H}}_0 - \mathbf{H}_0)\} = 0, \quad (\text{B.5})$$

where $\hat{\mathbf{H}}_0$ is constructed using $\mathbf{Y} - \mathbf{X}\tilde{\beta}$ in the estimators of the elements of $\hat{\Sigma}_{ww}$ defined in Appendix A and $\mathbf{H}_0 = \Sigma_{zz}^{-1} \Sigma_{ee}$. We set the covariance between $\tilde{\beta}$ and $\hat{\mathbf{H}}_\varphi$ equal to zero for all φ . Now

$$\begin{aligned} \hat{\mathbf{H}}_\varphi &= \hat{\Sigma}_{\varphi\varphi}^{-1} [\varphi \hat{\mathbf{D}}_{ee} + (1 - \varphi) \hat{\Sigma}_{ee}] \\ &= [\hat{\Sigma}_{ww} + \varphi \hat{\mathbf{D}}_{ee} + (1 - \varphi) \hat{\Sigma}_{ee}]^{-1} [\varphi \hat{\mathbf{D}}_{ee} + (1 - \varphi) \hat{\Sigma}_{ee}] \end{aligned}$$

and

$$\begin{aligned} \hat{\mathbf{H}}_\varphi - \mathbf{H}_\varphi &= \Sigma_{\varphi\varphi}^{-1} [\varphi (\hat{\mathbf{D}}_{ee} - \mathbf{D}_{ee}) + (1 - \varphi) (\hat{\Sigma}_{ee} - \Sigma_{ee})] \\ &\quad - \Sigma_{\varphi\varphi}^{-1} [\hat{\Sigma}_{ww} - \Sigma_{ww} + \varphi (\hat{\mathbf{D}}_{ee} - \mathbf{D}_{ee}) \\ &\quad + (1 - \varphi) (\hat{\Sigma}_{ee} - \Sigma_{ee})] \mathbf{H}_\varphi \\ &= \Sigma_{\varphi\varphi}^{-1} [\varphi (\hat{\mathbf{D}}_{ee} - \mathbf{D}_{ee} + (1 - \varphi) (\hat{\Sigma}_{ee} - \Sigma_{ee}))] \mathbf{G}_\varphi \\ &\quad - \Sigma_{\varphi\varphi}^{-1} [\hat{\Sigma}_{ww} - \Sigma_{ww}] \mathbf{H}_\varphi, \end{aligned}$$

where $\mathbf{G}_\varphi = \mathbf{I} - \mathbf{H}_\varphi$. The contribution of $\hat{\mathbf{D}}_{ee} - \mathbf{D}_{ee}$ to the variance of $\hat{\mathbf{H}}_\varphi$ is small relative to the contribution of $\hat{\Sigma}_{ee} - \Sigma_{ee}$. Therefore, we omit $\hat{\mathbf{D}}_{ee} - \mathbf{D}_{ee}$ in our variance approximation. Then the expectation

$$\begin{aligned} E\{(\mathbf{I} - \mathbf{H}_\varphi)' (\hat{\Sigma}_{ee} - \Sigma_{ee}) \Sigma_{\varphi\varphi}^{-1} \mathbf{z} \mathbf{z}' \Sigma_{\varphi\varphi}^{-1} \\ (\hat{\Sigma}_{ee} - \Sigma_{ee}) (\mathbf{I} - \mathbf{H}_\varphi)\} \\ = d_e^{-1} \mathbf{G}_\varphi' [\Sigma_{ee} (\text{tr}(\Lambda \Sigma_{ee})) + \Sigma_{ee} \Lambda \Sigma_{ee}] \mathbf{G}_\varphi, \quad (\text{B.6}) \end{aligned}$$

where $\Lambda = \Sigma_{\varphi\varphi}^{-1} \Sigma_{zz} \Sigma_{\varphi\varphi}^{-1}$, because \mathbf{z} is independent of $\hat{\Sigma}_{ee}$. We also omit the term $d_e^{-1} \mathbf{G}_\varphi' \Sigma_{ee} \Sigma_{\varphi\varphi}^{-1} \Sigma_{zz} \Sigma_{\varphi\varphi}^{-1} \Sigma_{ee} \mathbf{G}_\varphi$ in our variance approximation.

The expectation for the term containing $(\hat{\Sigma}_{ww} - \Sigma_{ww})$ is

$$E\{\mathbf{H}_\varphi' (\hat{\Sigma}_{ww} - \Sigma_{ww}) \Sigma_{\varphi\varphi}^{-1} \mathbf{z} \mathbf{z}' \Sigma_{\varphi\varphi}^{-1} (\hat{\Sigma}_{ww} - \Sigma_{ww}) \mathbf{H}_\varphi\},$$

where

$$\hat{\Sigma}_{ww} - \Sigma_{ww} = \begin{pmatrix} \mathbf{I}_{n1}(\hat{\sigma}_1^2 - \sigma_1^2) & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n2}(\hat{\sigma}_2^2 - \sigma_2^2) \end{pmatrix}.$$

Approximating the expectation by treating \mathbf{z} as independent of $\hat{\Sigma}_{ww}$, we obtain

$$\mathbf{H}_\varphi' \begin{pmatrix} \Lambda_{11} V\{\hat{\sigma}_1^2\} & \Lambda_{12} C\{\hat{\sigma}_1^2, \hat{\sigma}_2^2\} \\ \Lambda_{21} C\{\hat{\sigma}_1^2, \hat{\sigma}_2^2\} & \Lambda_{22} V\{\hat{\sigma}_2^2\} \end{pmatrix} \mathbf{H}_\varphi, \quad (\text{B.7})$$

where

$$\Lambda = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix} = \Sigma_{\varphi\varphi}^{-1} \Sigma_{zz} \Sigma_{\varphi\varphi}^{-1}.$$

The Taylor expansion of $\hat{\beta} - \beta$ is

$$\begin{aligned} \hat{\beta} - \beta &= (\mathbf{X}' \Sigma_{\varphi\varphi}^{-1} \mathbf{X})^{-1} \mathbf{X}' \hat{\Sigma}_{\varphi\varphi}^{-1} \mathbf{z} \\ &= (\mathbf{X}' \Sigma_{\varphi\varphi}^{-1} \mathbf{X})^{-1} \mathbf{X}' \Sigma_{\varphi\varphi}^{-1} \mathbf{z} \\ &\quad + (\mathbf{X}' \Sigma_{\varphi\varphi}^{-1} \mathbf{X})^{-1} \mathbf{X}' \Sigma_{\varphi\varphi}^{-1} (\hat{\Sigma}_{\varphi\varphi} - \Sigma_{\varphi\varphi}) \\ &\quad \times \Sigma_{\varphi\varphi}^{-1} \mathbf{X} (\mathbf{X}' \Sigma_{\varphi\varphi}^{-1} \mathbf{X})^{-1} \mathbf{X}' \Sigma_{\varphi\varphi}^{-1} \mathbf{z} \\ &\quad - (\mathbf{X}' \Sigma_{\varphi\varphi}^{-1} \mathbf{X})^{-1} \mathbf{X}' \Sigma_{\varphi\varphi}^{-1} \\ &\quad \times (\hat{\Sigma}_{\varphi\varphi} - \Sigma_{\varphi\varphi}) \Sigma_{\varphi\varphi}^{-1} \mathbf{z} + \text{Remainder}. \\ &= (\mathbf{X}' \Sigma_{\varphi\varphi}^{-1} \mathbf{X})^{-1} \mathbf{X}' \Sigma_{\varphi\varphi}^{-1} \mathbf{z} \\ &\quad + \mathbf{L} (\hat{\Sigma}_{\varphi\varphi} - \Sigma_{\varphi\varphi}) \Sigma_{\varphi\varphi}^{-1} \mathbf{Q} \Sigma_{\varphi\varphi}^{-1} \mathbf{z} \\ &\quad - \mathbf{L} (\hat{\Sigma}_{\varphi\varphi} - \Sigma_{\varphi\varphi}) \Sigma_{\varphi\varphi}^{-1} \mathbf{z} + \text{Remainder} \quad (\text{B.8}) \end{aligned}$$

where $\mathbf{Q} = \mathbf{X} (\mathbf{X}' \Sigma_{\varphi\varphi}^{-1} \mathbf{X})^{-1} \mathbf{X}'$ and $\mathbf{L} = (\mathbf{X}' \Sigma_{\varphi\varphi}^{-1} \mathbf{X})^{-1} \mathbf{X}' \Sigma_{\varphi\varphi}^{-1}$.

If $\Sigma_{\varphi\varphi} = \Sigma_{zz}$ and if $\hat{\Sigma}_{zz}$ is distributed as a multiple of a Wishart with d_e degrees of freedom, independent of \mathbf{z} , then

$$\begin{aligned} E\{\mathbf{L} (\hat{\Sigma}_{zz} - \Sigma_{zz}) \Sigma_{zz}^{-1} \mathbf{Q} \Sigma_{zz}^{-1} \mathbf{z} \mathbf{z}' \Sigma_{zz}^{-1} \mathbf{Q} \Sigma_{zz}^{-1} \\ \times (\hat{\Sigma}_{zz} - \Sigma_{zz}) \mathbf{L}'\} \\ = d_e^{-1} \mathbf{L} [\Sigma_{zz} \text{tr}(\Sigma_{zz}^{-1} \mathbf{Q}) + \mathbf{Q}] \mathbf{L}' \\ = d_e^{-1} (\mathbf{X}' \Sigma_{zz}^{-1} \mathbf{X})^{-1} (k+1). \end{aligned}$$

Using a similar approximation

$$\begin{aligned} E\{\mathbf{L} (\hat{\Sigma}_{zz} - \Sigma_{zz}) \Sigma_{zz}^{-1} \mathbf{X} \mathbf{L} \Sigma_{zz}^{-1} (\hat{\Sigma}_{zz} - \Sigma_{zz}) \mathbf{L}'\} \\ = E\{\mathbf{L} (\hat{\Sigma}_{zz} - \Sigma_{zz}) \Sigma_{zz}^{-1} \mathbf{X} \mathbf{L} \Sigma_{zz}^{-1} (\hat{\Sigma}_{zz} - \Sigma_{zz}) \mathbf{L}'\} \\ = E\{\mathbf{L} (\hat{\Sigma}_{zz} - \Sigma_{zz}) \Sigma_{zz}^{-1} \mathbf{Q} \Sigma_{zz}^{-1} (\hat{\Sigma}_{zz} - \Sigma_{zz}) \mathbf{L}'\} \\ = d_e^{-1} (\mathbf{X}' \Sigma_{zz}^{-1} \mathbf{X})^{-1} (k+1). \end{aligned}$$

On the basis of this result, we use the approximation

$$\hat{\beta} - \beta = (\mathbf{X}' \Sigma_{\varphi\varphi}^{-1} \mathbf{X})^{-1} \mathbf{X}' \Sigma_{\varphi\varphi}^{-1} \mathbf{z} - \mathbf{L} (\hat{\Sigma}_{\varphi\varphi} - \Sigma_{\varphi\varphi}) \Sigma_{\varphi\varphi}^{-1} \mathbf{z}.$$

We assume $\hat{\Sigma}_{ee}$ is a multiple of a Wishart matrix with d_e -degrees of freedom and approximate $\hat{\Sigma}_{\varphi\varphi} - \Sigma_{\varphi\varphi}$ with $(1-\varphi)(\hat{\Sigma}_{ee} - \Sigma_{ee})$. We have

$$\begin{aligned} & (1-\varphi)^2 E\left\{\mathbf{L}(\hat{\Sigma}_{ee} - \Sigma_{ee})\Sigma_{\varphi\varphi}^{-1}\Sigma_{zz}^{-1}(\hat{\Sigma}_{ee} - \Sigma_{ee})\mathbf{L}'\right\} \\ &= (1-\varphi)^2 d_e^{-1} \mathbf{L}\left[\Sigma_{ee} \text{tr}\left\{\Sigma_{\varphi\varphi}^{-1}\Sigma_{zz}^{-1}\Sigma_{\varphi\varphi}\Sigma_{ee}\right\}\right. \\ & \quad \left.+ \Sigma_{ee}\Sigma_{\varphi\varphi}^{-1}\Sigma_{zz}\Sigma_{\varphi\varphi}^{-1}\Sigma_{ee}\right]\mathbf{L}'. \end{aligned} \quad (\text{B.9})$$

The dominant term is that associated with the trace and we retain only that term in our approximation. Thus, an approximation to the variance of $\hat{\beta}$ is

$$\begin{aligned} \mathbf{V}_{\beta\beta} &= \mathbf{L}\Sigma_{zz}\mathbf{L}' \\ &+ d_e^{-1}(1-\varphi)^2 \text{tr}\left\{\Sigma_{\varphi\varphi}^{-1}\Sigma_{zz}^{-1}\Sigma_{\varphi\varphi}\Sigma_{ee}\right\} \mathbf{L}\Sigma_{ee}\mathbf{L}' \end{aligned} \quad (\text{B.10})$$

Combining results (B.6), (B.7), and (B.9), a crude estimator of the variance of the predictor (31) is

$$\begin{aligned} \hat{\mathbf{V}}\{\hat{\mathbf{y}}_\varphi\} &= \hat{\mathbf{H}}_\varphi' \hat{\Sigma}_{ww} \hat{\mathbf{H}}_\varphi + \hat{\mathbf{G}}_\varphi' \hat{\Sigma}_{ee} \hat{\mathbf{G}}_\varphi \\ &+ \hat{\mathbf{H}}_\varphi' \mathbf{X} \hat{\mathbf{V}}_{\beta\beta} \mathbf{X}' \hat{\mathbf{H}}_\varphi + \hat{\Gamma}_{44} + \hat{\Gamma}_{33}, \end{aligned} \quad (\text{B.11})$$

where

$$\hat{\mathbf{H}}_\varphi = \mathbf{I} - \hat{\mathbf{G}}_\varphi,$$

$$\begin{aligned} \hat{\mathbf{V}}_{\beta\beta} &= \hat{\mathbf{L}}_\varphi' \hat{\Sigma}_{zz} \hat{\mathbf{L}}_\varphi + d_e^{-1}(1-\varphi)^2 \\ &\times \text{tr}\left\{\hat{\Sigma}_{\varphi\varphi}^{-1} \hat{\Sigma}_{zz} \hat{\Sigma}_{\varphi\varphi}^{-1} \hat{\Sigma}_{ee}\right\} \hat{\mathbf{L}}_\varphi' \hat{\Sigma}_{ee} (1 + \delta_\varphi) \hat{\mathbf{L}}_\varphi', \end{aligned}$$

$$\hat{\mathbf{L}}_\varphi = \left(\mathbf{X}' \hat{\Sigma}_{\varphi\varphi}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}' \hat{\Sigma}_{\varphi\varphi}^{-1},$$

$$\hat{\Sigma}_{\varphi\varphi}^{-1} = \left(\hat{\Sigma}_{\varphi\varphi} + \delta_\varphi \hat{\Sigma}_{\varphi\varphi}\right)^{-1},$$

$$\hat{\Sigma}_{zz} = \hat{\Sigma}_{ww} + \hat{\Sigma}_{ee},$$

$$\delta_\varphi = \left[d_e - \text{tr}\left\{\hat{\Sigma}_{zz}^{-1} \hat{\Sigma}_{ee}\right\}\right]^{-1} \text{tr}\left\{\hat{\Sigma}_{zz}^{-1} \hat{\Sigma}_{ee}\right\},$$

$$\hat{\Gamma}_{44} = d_e^{-1}(1-\varphi)^2 \text{tr}\left\{\hat{\Sigma}_{\varphi\varphi}^{-1} \hat{\Sigma}_{zz} \hat{\Sigma}_{\varphi\varphi}^{-1} \hat{\Sigma}_{ee}\right\} \hat{\mathbf{G}}_\varphi' \hat{\Sigma}_{ee} \hat{\mathbf{G}}_\varphi,$$

$$\hat{\Gamma}_{33} = \hat{\mathbf{H}}_\varphi' \begin{pmatrix} \hat{\Lambda}_{11} \hat{V}\{\hat{\sigma}_1^2\} & \hat{\Lambda}_{12} \hat{C}\{\hat{\sigma}_1^2, \hat{\sigma}_2^2\} \\ \hat{\Lambda}_{21} \hat{C}\{\hat{\sigma}_1^2, \hat{\sigma}_2^2\} & \hat{\Lambda}_{22} \hat{V}\{\hat{\sigma}_2^2\} \end{pmatrix} \hat{\mathbf{H}}_\varphi,$$

$$\hat{\Lambda} = \begin{pmatrix} \hat{\Lambda}_{11} & \hat{\Lambda}_{12} \\ \hat{\Lambda}_{21} & \hat{\Lambda}_{22} \end{pmatrix} = \hat{\Sigma}_{\varphi\varphi}^{-1} \hat{\Sigma}_{zz} \hat{\Sigma}_{\varphi\varphi}^{-1},$$

$\hat{V}\{\hat{\sigma}_j^2\}$, $j=1,2$, is the estimated variance of $\hat{\sigma}_j^2$, and $\hat{C}\{\hat{\sigma}_1^2, \hat{\sigma}_2^2\}$ is the estimated covariance between $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$.

See Appendix A. The estimator of the variance of $\hat{\beta}$ contains an adjustment for the fact that $(\mathbf{X}' \hat{\Sigma}_{zz}^{-1} \mathbf{X})^{-1}$ is a biased estimator of $(\mathbf{X}' \Sigma_{zz}^{-1} \mathbf{X})^{-1}$.

REFERENCES

- BATTESE, G.E., HARTER, R.M., and FULLER, W.A. (1988). An error components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.
- CRESSIE, N. (1992). REML Estimation in Empirical Bayes smoothing of census undercount. *Survey Methodology*, 18, 75-94.
- EFRON, B., and MORRIS, C. (1972). Limiting the risk of Bayes and Empirical Bayes estimates - Part II: The Empirical Bayes case. *Journal of the American Statistical Association*, 67, 130-139.
- ERICKSEN, E.P., and KADANE, J.B. (1985). Estimating the population in a census year (with discussion). *Journal of the American Statistical Association*, 80, 98-131.
- ERICKSEN, E.P., KADANE, J.B., and TUKEY, J.W. (1989). Adjusting the 1980 Census of Population and Housing (with discussion). *Journal of the American Statistical Association*, 84, 927-944.
- FAY, R.E. (1987). Application of multivariate regression to small domain estimation. In *Small Area Statistics*, (Eds. R. Platek, J.N.K. Rao, C.-E. Särndal and M.P. Singh). New York: Wiley, 91-102.
- FAY, R.E. (1990). VPLX: Variance estimates for complex samples. *Proceedings of the Section on Survey Research Method, American Statistical Association*, 266-271.
- FAY, R.E. (1992). Inferences for Small Domain Estimates From the 1990 Post Enumeration Survey. Unpublished manuscript, U.S. Bureau of the Census.
- FAY, R.E., and HERRIOTT, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.
- FULLER, W.A., and HARTER, R.M. (1987). The multivariate components of variance model for small area estimation. In *Small Area Statistics*, (Eds., R. Platek, J.N.K. Rao, C.-E. Särndal and M.P. Singh), New York: Wiley, 103-123.
- GHOSH, M. (1992). Hierarchical and Empirical Bayes multivariate estimation. In *Current Issues in Statistical Inference: Essays in Honor of D. Basu*, (Eds. M. Ghosh and P.K. Pathak), IMS Lecture Notes Monograph Series, 17, 151-177.
- GHOSH, M., and RAO, J.N.K. (1994). Small area estimation: an appraisal. *Statistical Science*, 9, 55-93.
- HARVILLE, D.A. (1976). Extension of the Gauss-Markov Theorem to include estimation of random effects. *Annals of Statistics*, 4, 384-395.
- HENDERSON, C.R. (1950). Estimation of genetic parameters (Abstract). *Annals of Mathematical Statistics*, 21, 309-310.
- HOGAN, H. (1992). The 1990 Post Enumeration Survey: an overview. *The American Statistician*, 46, 261-269.

- HOGAN, H. (1993). The 1990 Post Enumeration Survey: operations and results. *Journal of the American Statistical Association*, 88, 1047-1060.
- HULTING, F.L., and HARVILLE, D.A. (1991). Some Bayesian and non-Bayesian procedures for the analysis of comparative experiments and small area estimation: computational aspects, frequentist properties, and relationships. *Journal of the American Statistical Association*, 86, 557-568.
- ISAKI, C.T., HUANG, E.T., and TSAY, J.H. (1991). Smoothing adjustment factors from the 1990 Post Enumeration Survey. *Proceedings of the Social Statistics Section, American Statistical Association*, 338-343.
- KACKAR, R.N., and HARVILLE, D.A. (1984). Approximations for standard errors of estimators for fixed and random effects in mixed models. *Journal of the American Statistical Association*, 79, 853-862.
- MORRIS, C. (1983). Parametric Empirical Bayes inference: theory and applications (with discussions). *Journal of the American Statistical Association*, 78, 47-65.
- PEIXOTO, J.L., and HARVILLE, D.A. (1986). Comparisons of alternative predictors under the balanced one-way random model. *Journal of the American Statistical Association*, 81, 431-436.
- PRASAD, N.G.N., and RAO, J.N.K. (1990). The estimation of mean squared errors of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.
- ROBINSON, G.K. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science*, 6, 15-51.
- SEARLE, S.R. (1971). *Linear Models*. New York: Wiley.
- SINGH, M.P., GAMBINO, J., and MANTEL, H.J. (1994). Issues and strategies for small area data. *Survey Methodology*, 20, 3-14.
- U.S. BUREAU OF THE CENSUS. (1988). The Coverage of Population in the 1980 Census, Evaluation and Research Program, PHC(E)-4.

Census Coverage Error: A Demographic Evaluation

RÉJEAN LACHAPELLE and DON KERR¹

ABSTRACT

The 1996 Canadian Census is adjusted for coverage error as estimated primarily through the Reverse Record Check (RRC). In this paper, we will show how there is a wealth of additional information from the 1996 Reverse Record Check of direct value to population estimation. Beyond its ability to estimate coverage error, it is possible to extend the Reverse Record Check classification results to obtain an alternative estimate of demographic growth – potentially decomposed by component. This added feature of the Reverse Record Check provides promise in the evaluation of estimated census coverage error as well as insight as to possible problems in the estimation of selected components in the population estimates program.

KEY WORDS: Census coverage error; Population estimates; Reverse record check.

1. INTRODUCTION

The Reverse Record Check (RRC), in various forms, has been used by Statistics Canada since the 1960's to estimate coverage error in the Canadian Census (Fellegi 1969; Brackstone and Gosselin 1973; Gosselin 1976; Burgess 1988; Carter 1990; Royce, Germain, Julien, Dick, Switzer and Allard 1994, Statistics Canada 1999). Using the Reverse Record Check, Statistics Canada has produced a long time series of population estimates, from 1971 through to the present, fully adjusted for census undercount. The current paper will demonstrate how there is additional information in the Reverse Record Check, which from a demographic perspective, can be exploited for the purposes of population estimation.

The demographic statistics program at Statistics Canada uses information from vital statistics, the most recent census, and various administrative sources in generating highly accurate and up to date population estimates. Information on births, deaths, immigration, emigration, among other demographic components, can be used to estimate population growth since the previous census. With each quinquennial census, a cycle ends and the accuracy of these estimates are put to the test (Romaniuc 1988). Systematic comparisons can be made between these estimates of growth and estimated growth as implied by comparing subsequent censuses (after adjustment for census coverage error).

The resultant difference (conventionally referred to as the error of closure of the intercensal population estimates) has a far from obvious interpretation. While a large error of closure is suggestive of problems in the population estimates, its specific nature is far from obvious (as to which demographic components are specifically responsible for the error). Furthermore, a honest appraisal of this closure error might suggest not only problems in the population estimates, but also potential problems in census coverage

studies themselves (at the beginning and/or end of the intercensal period).

The current paper will demonstrate how an alternative estimate of demographic growth is possible, as based explicitly on the RRC classification results. Additional information is available, which assists greatly in the interpretation and decomposition of this closure error. Three alternative estimates of demographic growth for the intercensal period will be presented in the following section, including growth as estimated as part of the regular program of population estimates, implicit growth as obtained in comparing consecutive censuses, and growth as based explicitly on RRC classification results. Section 3 demonstrates how this RRC based estimate of growth can assist in the decomposition and interpretation of closure error, providing evidence of (i) bias in selected components of the population estimates, and (ii) possible problems in the RRC results. Section 4 presents the results from this decomposition, followed by a brief discussion of its implications for both census coverage error measurement and the population estimates program.

2. ALTERNATIVE ESTIMATES OF DEMOGRAPHIC GROWTH

2.1. Administrative Record Based Estimates of Growth: Post-Censal Estimates

Statistics Canada's regular program of population estimates involves the continuous registration and estimation of demographic events, as based on vital statistics and various administrative data sets. These events are added or subtracted from the population documented in the previous census (component method). In estimating a province's population on Census day 1996 (P_{est96}):

¹ Réjean Lachapelle, Demography Division, Main Building, Tunney's Pasture, Statistics Canada, Ottawa, Ontario, K1A 0T6; Don Kerr, Department of Sociology, University of Western Ontario, London, Ontario, N6A 5C2.

$$P_{\text{est}96} = P_{91} + B_{91-96} - D_{91-96} + I_{91-96} - E_{91-96} + \Delta \text{NPR}_{91-96} + \text{NM}_{91-96} \quad (1)$$

The baseline population (P_{91}) for this estimate builds on the 1991 Census after adjustment for all forms of coverage error, including net census undercount as measured through the 1991 RRC. The postcensal estimate can be obtained by adding or subtracting from this baseline the number of births between censuses (B_{91-96}), the number of deaths (D_{91-96}), immigrants (I_{91-96}), emigrants (E_{91-96}), net interprovincial migration (NM_{91-96}), and the net gain or loss of nonpermanent residents ($\Delta \text{NPR}_{91-96}$).

Non-permanent residents (NPRs) are persons with legal temporary status in Canada (e.g., persons holding student or employment authorizations, minister's permits, refugee claimants, as well as their non-Canadian born dependents). Unlike with interprovincial migration, net gain or net loss of NPRs is not estimated through "flow" data on the ongoing in and out-flows of non-permanent residents, but rather estimated by comparing over time "stock" data on the total number of non-permanent residents living in the country. Further details of methodology, data sources and data quality issues can be obtained from the quarterly and annual releases of the population estimates program (Statistics Canada 1999; 2000).

2.2. Implicit Estimate of Growth

An implicit estimate of growth can be derived using the 1991 and 1996 Censuses, with both censuses adjusted for net undercount. With the exception of a small number of refusal Indian reserves, whose population figures are estimated independently, gross undercoverage was estimated entirely through the RRC in 1996, whereas gross overcoverage was a combined estimate from three studies (the RRC, the Collective Dwellings Study and the Automated Match Study). In 1991, the RRC was used only in the estimation of gross undercoverage, whereas gross overcoverage was estimated through a smaller study, the Private Dwelling Study, in combination with the 1991 Collective Dwelling and Automated Match studies. In addition, persons missed on refusal Indian reserves were estimated as part of the 1991 Reverse Record Check.

In the early evaluation of the 1996 coverage studies, the implicit growth obtained with the above adjustments was considered unrealistic. It has since been established that part of the 1991 estimate of net undercount was in error, and would have in reality been lower had selected methodological enhancements been introduced as in 1996 (Tourigny, Clark and Provost 1998). It has been shown that (i) a number of persons initially classified as missed in 1991 was too high due to misclassification, and (ii) the 1991 estimate of "overcount" was too low. As a result, 1991 estimates of undercount and overcount have been revised to reflect the impact of these methodological changes (${}^{\text{rev}}U_{91}$, ${}^{\text{rev}}O_{91}$). In addition, for reasons of consistency with

1996, separate modeled estimates of refusal Indian reserves (independent of the RRC) have been added to the Census in 1991.

More specifically, implicit growth (Δ^I) is obtained as:

$$\Delta^I = P_{96} - P_{91} = \{P_{96}^c + U_{96} - O_{96} + \text{IR}_{96M} - \text{IR}_{\text{RRC}96}\} - \{P_{91}^c + {}^{\text{rev}}U_{91} - {}^{\text{rev}}O_{91} + \text{IR}_{91M} - \text{IR}_{\text{RRC}91}\} \quad (2)$$

where final population figures (P_{96} , P_{91}) are obtained using previously published census figures (P_{96}^c , P_{91}^c) adjusted for undercoverage (U_{96} , ${}^{\text{rev}}U_{91}$) and gross overcoverage (O_{96} , ${}^{\text{rev}}O_{91}$). In adding independently modeled estimates of refusal Indian reserves (IR_{96M} , IR_{91M}), it is necessary to remove that portion of the RRC estimate of gross undercoverage that corresponds to these reserves ($\text{IR}_{\text{RRC}96}$, $\text{IR}_{\text{RRC}91}$). The results presented in the current paper take these changes into consideration.

2.3.1. RRC Based Estimates of Growth

The Reverse Record Check (RRC) is a record linkage and matching procedure that attempts to trace all persons in its sample, interview them to obtain a census day address, and match their records to individual census documents. This involves the construction of a sample intended to represent the same target population as the census being evaluated. This sampling frame, obtained in a manner that is totally independent of the census being evaluated, is constructed using the previous census, birth registrations over the intercensal period, administrative lists of inter-censal immigrants, and an up-to-date listings of non-permanent residents. Persons missed in the previous census are represented by a sample of cases classified as "missed" in the previous RRC, in the absence of a complete list of such persons.

By working with this sample, the RRC targets all persons who could have potentially been part of the 1996 Census universe. Except for a very small sub-population of returning emigrants (Canadian citizens and landed immigrants who were abroad during the previous census), the RRC sample is complete and fully representative. The subsequent classification (missed, enumerated, emigrated, abroad, deceased or out of scope) is applied in the estimation of "missed" in the current census. At the same time, this classification also holds the potential for further inferences, i.e., an additional estimate of demographic growth for the intercensal period.

To estimate demographic growth using the RRC, it is useful to consider the following two equations. In the first equation, the target population of the 1991 Census (P_{91}^T) is expressed in terms of all potential classification outcomes in 1996. In the second equation, it is possible to move in the opposite direction – by expressing the 1996 census target population (P_{96}^T) in terms of all possible statuses in 1991 (or in the case of births and immigrants, the intercensal period).

$$P_{91}^T = {}^{91}PP_{96} + {}^{91}NP_{96} + {}^{91}NP C_{96PP} + {}^{91}PP D_{96} + {}^{91}NP D_{96} + {}^{91}PP E_{96FR} + {}^{91}NP E_{96FR} + {}^{91}NP E_{96EX} \quad (3)$$

$$P_{96}^T = {}^{91}PP_{96} + {}^{91}NP_{96} + {}^{91}NP C_{96PP} + {}^{91-96}B_{96} + {}^{91}EX I_{96PP} + {}^{91}EX I_{96NP} + {}^{91}FR RE_{96PP} \quad (4)$$

where:

- ${}^{91}PP_{96}$ - Canadian citizens and landed immigrants in Canada in 1991, also targeted by the 1996 census
- ${}^{91}NP_{96}$ - NPRs in Canada in 1991, also targeted by the 1996 census as NPRs
- ${}^{91}NP C_{96PP}$ - NPRs in Canada in 1991 who became landed immigrants over the intercensal period
- ${}^{91}PP D_{96}$ - Canadian citizens and landed immigrants in Canada in 1991, who died over the intercensal period
- ${}^{91}NP D_{96}$ - NPRs in Canada in 1991, who died over the intercensal period
- FR - persons with the right to live permanently in Canada (citizens and landed immigrants) that are not in the designated census target population
- ${}^{91}PP E_{96FR}$ - Canadian citizens and landed immigrants in Canada in 1991, who are outside the 1996 Census target population
- ${}^{91}NP E_{96FR}$ - NPRs in Canada in 1991, who became landed immigrants or citizens, and are outside the 1996 census target population
- EX - persons who have never been citizens or landed immigrants, and are not in the designated census target population
- ${}^{91}NP E_{96EX}$ - NPRs in Canada in 1991, who did not become landed immigrants, and are outside the 1996 census target population
- ${}^{91-96}B_{96}$ - births over the 1991-1996 period, and in the 1996 census target population
- ${}^{91}EX I_{96NP}$ - persons not in Canada in 1991, who arrived over the intercensal period, and are NPRs in the 1996 census target population
- ${}^{91}EX I_{96PP}$ - immigrants who landed over the intercensal period, and are in the 1996 Census target population
- ${}^{91}FR RE_{96PP}$ - returning emigrants, *i.e.*, Canadian citizens and landed immigrants outside the census universe in 1991, and in the 1996 Census universe

An estimate of growth (Δ^{RRC}) can be obtained by subtracting the former equation from the latter:

$$\Delta^{RRC} = {}^{91-96}B_{96} + {}^{91}EX I_{96PP} + {}^{91}EX I_{96NP} - {}^{91}PP D_{96} - {}^{91}NP D_{96} - {}^{91}PP E_{96FR} - {}^{91}NP E_{96FR} - {}^{91}NP E_{96EX} + {}^{91}FR RE_{96PP}. \quad (5)$$

With the previously introduced sampling frames and classification outcomes, all terms (with the exception of the last term: returning emigrants) can be directly estimated from the 1996 RRC itself. The census target population in 1991 can be approximated through the sample drawn from the census and missed frames – with the identification of relevant classification outcomes. The census target population in 1996 can be approximated through all persons classified as either enumerated or missed in 1996. The final term (*i.e.*, returning emigrants) can be obtained independent of the RRC using the 1996 Census 5-year mobility variable, in identifying all persons outside the country five years ago (excluding recent immigrants and NPRs). It is possible to express this same RRC based estimate of demographic growth at the provincial level by incorporating an estimate of interprovincial migration. As the RRC relied on Health Care Files in Canada's two northern territories (the Yukon and NWT) with administrative lists of addresses current to census being evaluated, this estimate of growth is not possible for the relatively small populations living in Canada's far north.

A minor problem in the RRC design persists that potentially introduces a slight bias into its classification results. Unfortunately it is not possible to identify all NPRs in the RRC sample, with the potential for an unknown amount of frame overlap (*i.e.*, between the census, NPR and immigrant frames). As NPRs in the census can only be identified through the census long form (which is distributed to about 20% of all households), it is possible that some NPRs living in Canada in 1991, selected in the census frame, were also selected in either the immigrant or NPR frames (without being identified as such). While the RRC attempts to adjust for this overlap by identifying all such persons in the immigrant and NPR frames, an unknown bias exists to the extent that this is unsuccessful. This difficulty in identifying overlap leaves the potential of too many immigrants and/or NPRs in the sample, or too few, if too many persons are removed from the aforementioned frames. The latter outcome can subsequently deflate the estimate of demographic growth, gross undercoverage (among other classification outcomes), whereas the former has the opposite outcome.

2.3.2. RRC based Estimate of Growth: A More Detailed Decomposition

While both the postcensal and RRC based estimates of demographic growth should be highly comparable, the specific terms within each are not meant to be directly

equivalent. For example, births in the postcensal estimates denote all intercensal births occurring to a population – irrespective of whether such births move or die – whereas births in the discrete equation denote all births occurring yet still in Canada at the end of the intercensal period. With this in mind, it is possible to expand on the RRC based equation, to derive terms that are more comparable to those used in the postcensal estimates. The RRC based estimate of demographic growth can then be used in the evaluation of the components of demographic growth that enter into the component method.

To expand on this equation, it is useful to begin with births, again expressed in terms of possible RRC classification outcomes. As previously indicated, the birth term as included in equation (5) is only part of all births occurring over the intercensal period. More comprehensively, all births can be expressed as:

$$B^{91-96} = {}^{91-96}B_{96} + {}^{91-96B}D_{96} + {}^{91-96B}E_{96FR} \quad (6)$$

where:

$$\begin{aligned} B^{91-96} &= \text{all intercensal births} \\ {}^{91-96}B_{96} &= \text{all intercensal births ultimately classified as either enumerated or missed in 1996} \\ {}^{91-96B}D_{96} &= \text{deaths of intercensal births} \\ {}^{91-96B}E_{96FR} &= \text{persons outside target population in 1996 yet born in Canada over the intercensal period} \end{aligned}$$

Similarly, all immigrants can be expressed as:

$$I^{91-96} = {}^{91EX}I_{96PP} + {}^{91NP}C_{96PP} + {}^{91-96I}D_{96} + {}^{91-96I}E_{96FR} \quad (7)$$

where:

$$\begin{aligned} {}^{91EX}I_{96PP} &= \text{intercensal immigrants ultimately classified as either enumerated or missed in 1996} \\ {}^{91NP}C_{96PP} &= \text{all NPRs in 1991 who obtain landed immigrant status and are ultimately classified as either enumerated or missed in 1996} \\ {}^{91-96I}D_{96} &= \text{deaths occurring to landed immigrants over the intercensal period} \\ {}^{91-96I}E_{96FR} &= \text{emigrants among intercensal immigrants (irrespective of whether or not they were living in Canada as NPRs in 1991)} \end{aligned}$$

In combining equations 5, 6 and 7, demographic growth can be restated as:

$$\begin{aligned} P_{96}^T - P_{91}^T = & B^{91-96} - {}^{91PP}D_{96} - {}^{91NP}D_{96} - {}^{91-96B}D_{96} - \\ & {}^{91-96I}D_{96} + I^{91-96} + {}^{91FR}RE_{96PP} - {}^{91NP}C_{96PP} - \\ & {}^{91PP}E_{96FR} - {}^{91NP}E_{96FR} - {}^{91-96B}E_{96FR} - {}^{91-96I}E_{96FR} - \\ & {}^{91NP}E_{96EX} - {}^{91EX}I_{96NP} \end{aligned} \quad (8)$$

Given that the final term of (8) is equivalent to:

$${}^{91EX}I_{96NP} = NP_{96} - NP_{91} + {}^{91NP}D_{96} + {}^{91NP}E_{96EX} + {}^{91NP}C_{96PP} + {}^{91NP}E_{96FR} \quad (9)$$

It follows that:

$$\begin{aligned} P_{96}^T - P_{91}^T = & B^{91-96} - {}^{91PP}D_{96} - {}^{91NP}D_{96} - {}^{91-96B}D_{96} - {}^{91-96I}D_{96} \\ & + I^{91-96} + {}^{91FR}RE_{96PP} - {}^{91NP}C_{96PP} - \\ & {}^{91PP}E_{96FR} - {}^{91NP}E_{96FR} - {}^{91-96B}E_{96FR} - \\ & {}^{91-96I}E_{96FR} - {}^{91NP}E_{96EX} + NP_{96} - (NP_{91} - {}^{91NP}D_{96} - \\ & {}^{91NP}E_{96EX} - {}^{91NP}C_{96PP} - {}^{91NP}E_{96FR}) \end{aligned} \quad (10)$$

or:

$$\begin{aligned} P_{96}^T - P_{91}^T = & (B^{91-96} - ({}^{91PP}D_{96} - {}^{91-96B}D_{96} - {}^{91-96I}D_{96})) + (I^{91-96} - \\ & ({}^{91PP}E_{96FR} + {}^{91-96B}E_{96FR} + {}^{91-96I}E_{96FR} - {}^{91FR}RE_{96PP})) + \\ & (NP_{96} - NP_{91}) \end{aligned} \quad (11)$$

This expanded version of equation (5) provides a breakdown of demographic growth at the national level, and allows for more meaningful comparisons with components estimated through administrative records. All terms, except for ${}^{91FR}RE_{96PP}$ and NP_{91} can be obtained directly from the 1996 RRC. The aforementioned hole in the RRC sampling frame requires an independent estimate of returning emigrants whereas the nature of the sample frame for NPRs explains the absence of the latter term. Rather than a listing of all NPRs to enter Canada over the intercensal period (as was the case with immigrants), the RRC relies on the most up to date administrative listing of NPRs in the establishment of its sampling frame (with no information on the number of NPRs living in Canada in 1991).

Postcensal estimates document demographic growth through the “continuous” registration and estimation of demographic events over time. The RRC estimates growth via information on the status of individuals as identified on at least two “discrete” dates (at the beginning and end of the intercensal period). Irrespective of this minor conceptual distinction between “continuous” versus “discrete” estimation, each term of equation 11 (within each set of parenthesis) roughly corresponds to a separate component as documented using administrative records. The first term identifies all intercensal births (*i.e.*, the weighted sum of the birth frame), the second term includes deaths (classification results across the birth frame, the missed frame, the census frame and immigrant frame), the third term includes all

immigrants (*i.e.*, the weighted sum of the immigrant frame), the fourth term includes emigrants (classification results across the birth frame, the immigrant frame, the missed frame and census frames, as well as the returning emigrant component), and the fifth term corresponds to net gain or loss of NPRs. As the number of NPRs living in Canada in 1991 is not available in the 1996 RRC, for current purposes, this latter term is obtained using the 1991 census count, after adjustments for undercoverage. Again, it is possible to express this equation at the provincial level.

With equation (11), a detailed evaluation of the postcensal estimation program is possible. For example, if differences persist between RRC based estimates and postcensal estimates, it is possible to determine how much of the difference in estimated growth can be traced back to differences in migration (typically estimated with some difficulty in the postcensal estimates program) and how much can be traced to differences in natural increase. Briefly, Table 1 includes all of the aforementioned estimates of growth, including implicit growth, the growth as based on administrative records, and the two alternate estimates of growth as based on the RRC (simplified and expanded equations). Slight differences exist between the simplified and expanded equations – yet not nearly of the same size as with the other estimates (implicit, postcensal). In explanation of the differences between the two RRC based estimates, the simplified equation does not require the same detailed classification as with the expanded equation, is not biased to the same extent by the aforementioned problem of frame overlap, and does not rely on the 1991 census count of NPRs. The differences observed with the remaining estimates are the focus of the current decomposition.

Table 1
Alternate Estimates of Growth, 1991-1996, Canada and
Provinces/Territories

	Implicit Growth	Population Estimates Administrative records	RRC simplified	RRC expanded
NFLD.	-17,997	-9,263	-17,897	-17,751
P.E.I.	5,404	5,483	2,568	1,583
N.S.	15,781	24,271	17,075	16,860
N.B.	7,714	13,097	12,017	11,276
QUE.	206,307	300,849	261,357	252,014
ONT.	659,349	766,568	668,443	655,572
MAN.	23,682	24,981	7,377	6,288
SASK.	15,953	11,098	11,524	9,312
ALTA.	186,594	186,986	151,944	159,907
B.C.	505,025	466,611	465,864	472,342
YUKON	3,085	2,329	N/A	N/A
N.W.T.	6,837	5,864	N/A	N/A
Provinces (excl terr)	1,607,771	1,790,681	1,580,273	1,567,404
Canada	1,617,693	1,798,874	N/A	N/A

3. A DECOMPOSITION OF CLOSURE ERROR

Implicit growth for the 1991-96 period is obtained only after all adjustments have been made to the censuses for coverage error, including revised 1991 figures on gross undercount and overcount and refinements for refusal Indian reserves. Alternatively, the RRC based estimate of growth (simplified version) is obtained in working with approximations of the 1991 and 1996 target populations, *i.e.*, the census and missed frames of the 1996 RRC and all persons classified as either missed or enumerated in this study. For this reason, there are minor differences between the two estimates that need to be more clearly identified in a full decomposition of closure error. In this context, it is useful to express implicit growth obtained with final population figures in terms of these approximations (sampling frames and classification outcomes). In a similar manner, as the error of closure is the difference between implicit growth and the growth associated with the postcensal estimates, the error of closure can also be expressed in terms of these approximations.

To simplify the presentation, let δ represent all possible negative growth terms in equation (5) and η as all possible positive growth terms:

$$\delta = ({}^{91}PP_{96} + {}^{91}NP_{96} + {}^{91}PP_{96FR} + {}^{91}NP_{96FR} + {}^{91}NP_{96EX}) \quad (12)$$

$$\eta = ({}^{91-96}B_{96} + {}^{91EX}I_{96PP} + {}^{91EX}I_{96NP} + {}^{91FR}RE_{96PP}) \quad (13)$$

The population enumerated in both censuses can be represented as:

$${}^{91}P_{96} = ({}^{91}PP_{96} + {}^{91}NP_{96} + {}^{91}NP_{C_{96PP}}) \quad (14)$$

Since the final population figures (P_{91} , P_{96}) used in the estimation of implicit growth involve separate modeled estimates for refusal Indian reserves, it is useful to restate the RRC based estimate of growth after specifically delineating such reserves. In designating persons living in refusal reserves in 1996 that were in the target population in 1991 as ${}^{91}IR_{96}$, the growth of these reserves through either migration or birth as estimated by the RRC by η_{IR} , and redefining ${}^{91}P_{96}$ to exclude all persons associated with these two terms, it is possible to return to equations (3)-(5) as:

$$P_{91}^T = {}^{91}P_{96} + {}^{91}IR_{96} + \delta \quad (15)$$

$$P_{96}^T = {}^{91}P_{96} + {}^{91}IR_{96} + \eta_{IR} + \eta \quad (16)$$

$$\Delta^{RRC} = P_{96} - P_{91} = \eta + \eta_{IR} - \delta \quad (17)$$

In expressing implicit growth in terms of the RRC sampling frames and classification outcomes, it is useful to build on the RRC estimate of growth (equation 17) in defining total growth beginning with P_{91} rather than P_{91}^T . In recognition that the final population estimate (P_{91}) is equivalent to the census and missed frames (P_{91}^T) minus overcoverage (${}^{\text{rev}}O_{91}$) plus refinements for refusal Indian reserves ($IR_{91M} - IR_{RRC91}$), it follows:

$$P_{96}^T - P_{91} = \eta + n_{IR} - \delta + {}^{\text{rev}}O_{91} + (IR_{RRC91} - IR_{91M}). \quad (18)$$

On the other hand, this target population (P_{96}^T) can also be expressed as:

$$P_{96}^T = EN_{96} + U_{96} + {}^{91FR}RE_{96PP} \quad (19)$$

where EN_{96} is an estimate of the number of persons enumerated in 1996. In recalling from equation 2 that:

$$P_{96} = P_{96}^c + U_{96} - O_{96} + (IR_{96M} - IR_{RRC96}) \quad (20)$$

implicit growth (Δ^I) can be expressed in terms of the RRC based estimates of growth, as:

$$\begin{aligned} \Delta^I &= P_{96} - P_{91} = (P_{96} - P_{96}^T) + (P_{96}^T - P_{91}) = \\ &= \{(\eta - \delta)\} + \{\eta_{IR} - (IR_{91M} - IR_{RRC91}) + (IR_{96M} - IR_{RRC96})\} + \\ &= \{(P_{96}^c - EN_{96} - {}^{91FR}RE_{96PP} + {}^{\text{rev}}O_{91} - O_{96})\}. \end{aligned} \quad (21)$$

Implicit growth (Δ^I) can be defined as the sum of (i) a RRC based estimate of growth (excluding refusal Indian reserves), (ii) a second term depending on the decision to estimate the refusal Indian reserves by an independent model, and (iii) a third term that involves a comparison of the RRC based estimate of enumerated and the number of persons actually enumerated in the 1996 census.

This latter term (the difference on enumerated) has an interesting interpretation, and is considered fundamental to the evaluation of the RRC (Tourigny, Bureau and Clark 1998; Royce 1993). Significant differences with this term can be read as implying either sampling errors and/or possible biases, as either classification error and/or problems in sample selection. To make this comparison meaningful, 1996 overcoverage and an estimate of returning emigrants are removed from the census counts - as neither can be included in the estimate of enumerated. Similarly, since the RRC selects part of its sample from the previous census, it inevitably carries forward some overcoverage inherent in its weights - which must subsequently be removed from its estimate of enumerated. These adjustments are included in the third term (the third set of brackets) in equation 21.

While the estimate of enumerated is inflated by the weights associated with overcoverage in the 1991 Census

frame, only a portion is directly associated with this estimate - with the remainder spread across the other classification outcomes. Consequently, all classification results in the aforementioned equations are also slightly overstated. For the purposes of the current decomposition, this minor distinction is ignored. This is another reason, albeit of minor impact, why the RRC-based estimate of growth is different from the implicit estimate, as the latter is not biased by this overcoverage.

From the above, the error of closure is equivalent to:

$$\begin{aligned} \Delta_{91-96}^D - \Delta_{91-96}^I &= \\ &= [\Delta_{91-96}^D - \{(\eta - \delta)\} - \\ &\quad \{\eta_{IR} - (IR_{91M} - IR_{RRC91}) + (IR_{96M} - IR_{RRC96})\}] - \\ &= \{(P_{96}^c - EN_{96} - {}^{91FR}RE_{96PP} + {}^{\text{rev}}O_{91} - O_{96})\}. \end{aligned} \quad (22)$$

In the decomposition of closure error, the first term inside brackets [] highlights the difference between the postcensal estimate of growth and the combined RRC estimate of growth (including refusal reserves, after refinements for modeled estimates). The second term (the difference on enumerated) provides evidence as to possible difficulties in the coverage studies. Theoretically, with the absence of sampling and non-sampling error in the RRC, this latter term should be negligible.

4.1. Decomposition Results: Closure Error

Table 2 presents closure error after finalizing both the 1991 and 1996 estimates of population. By adding net undercount to the 1996 published Census figures, along with independent estimates of refusal Indian reserves, Canada's 1996 Census day population, adjusted for coverage error is estimated at 29,619,539. This figure is appreciably lower than the Census day estimate as generated through the postcensal estimates program of 29,800,720. The difference between the two figures - which is equivalent to the aforementioned difference between implicit growth and growth as based on administrative records - was higher than anticipated given past experience, at 181,181 (or .61% of the 1991 Census Day population).

Across provinces/territories, closure error is found to be particularly pronounced in Newfoundland (1.56%), in Canada's north (at -2.38% in Yukon and -1.44% in the NWT), and somewhat surprisingly, in its three largest provinces (as 1.30% in Quebec, .97% in Ontario and -.99% in British Columbia). Regionally, closure errors larger than the national average are observed across eastern and central Canada (except for P.E.I.) while the western provinces have closure errors lower than the national one. It is specifically these errors that the current decomposition seek to evaluate and explain.

Table 3 presents the results from this decomposition, with closure error decomposed into (i) the difference

between the estimate of growth based on administrative records and the RRC based estimate (simplified version), and (ii) the difference on enumerated. Also included is the sampling error associated with the RRC estimates.

4.2. Comparisons between Estimates of Growth

Across all provinces (with the exception of Saskatchewan), growth estimated on the basis of administrative records is higher than the RRC based estimate. At

the national level (excluding the territories), this discrepancy on growth (210,408) appears far more important in explaining closure error than the discrepancy on enumerated (-27,498). While for many provinces the difference on growth fell well within expectations in light of sampling error, selected provinces require further explanation. For example, the difference in growth in Ontario is large (98,125), which is almost one half the difference observed

Table 2
Coverage Study Results, Relative to Population Estimate (1996 - Census Day)

	{1}	{2}	{3}	{4=1+2+3}	{5}	{6=5-4}	{7=6/4*100}
	1996 census count with random additions	1996 net undercount	Indian Reserves	1996 Census RRC adjusted	1996 estimate post-censal (i)	Error of closure	Error of closure (%)
NFLD.	551,792	9,424	0	561,216	569,950	8,734	1.56
P.E.I.	134,557	1,149	175	135,881	135,960	79	0.06
N.S.	909,282	20,821	0	930,103	938,593	8,490	0.91
N.B.	738,133	14,225	518	752,876	758,259	5,383	0.71
QUE.	7,138,795	116,750	12,427	7,267,972	7,362,514	94,542	1.30
ONT.	10,753,573	301,368	20,849	11,075,790	11,183,050	107,260	0.97
MAN.	1,113,898	18,881	315	1,133,094	1,134,393	1,299	0.11
SASK.	990,237	28,051	586	1,018,874	1,014,019	-4,855	-0.48
ALTA.	2,696,826	66,327	11,287	2,774,440	2,774,832	392	0.01
B.C.	3,724,500	142,443	3,136	3,870,079	3,831,665	-38,414	-0.99
YUKON	30,766	1,022	0	31,788	31,032	-756	-2.38
N.W.T.	64,402	3,024	0	67,426	66,453	-973	-1.44
Canada	28,846,761	723,485	49,293	29,619,539	29,800,720	181,181	0.61

(i) Post-Censal Estimates for May 14th, obtained with final components for intercensal estimates. Final Estimates (Sept. 24th, 1998) of Net Undercount, 1991 an 1996.

Table 3
Decomposition of Closure Error

Province/Territory	Error of Closure	Difference between Dem. and RRC Estimates of Growth	S.E. of estimates	Difference on enumerated	S.E. of estimates
NFLD.	8,734	8,634	4,889	100	5,176
P.E.I.	79	2,915	2,425	-2,836	2,462
N.S.	8,490	7,196	9,011	1,294	9,455
N.B.	5,383	1,080	7,793	4,303	7,918
QUE.	94,542	39,492	25,493	55,050	29,310
ONT.	107,260	98,125	41,212	9,135	51,300
MAN.	1,299	17,604	10,108	-16,305	10,370
SASK.	-4,855	-426	9,187	-4,429	10,200
ALTA.	392	35,042	19,067	-34,650	21,618
B.C.	-38,414	747	20,518	-39,161	22,996
YUKON	-756	N/A	N/A	-108	270
N.W.T.	-973	N/A	N/A	-284	464
Canada without Territories	182,910	210,408	43,951	-27,498	58,724
Canada	181,181	N/A	N/A	-27,890	58,762

Table 4

Estimated Components (1991-1996) as Compiled by Demography Division and RRC Discrete (detailed) Measurement

	NFLD	PEI	NS	NB	QUE	ONT	MAN	SASK	ALB	BC	CANADA (without terr)
Births											
Demography	31,748	8,803	55,994	44,444	453,556	730,520	81,485	70,382	199,484	229,511	1,905,927
RRC	31,779	8,782	55,984	44,444	454,332	729,744	81,485	70,382	199,484	229,511	1,905,927
Difference	-31	22	10	0	-776	776	0	0	0	0	0
Deaths											
Demography	-19,286	-5,692	-37,677	-28,567	-252,628	-376,760	-45,858	-40,652	-75,798	-126,935	-1,009,853
RRC	-18,530	-6,913	-43,820	-29,354	-273,617	-400,047	-56,108	-40,143	-74,640	-138,433	-1,081,605
Difference	-756	1,221	6,143	787	20,989	23,287	10,250	-509	-1,158	11,498	71,752
Immigration											
Demography	3,411	771	14,489	3,359	189,905	618,869	22,004	11,282	84,130	213,506	1,161,726
RRC	3,538	820	14,058	3,614	189,905	618,870	22,129	11,157	84,130	216,892	1,165,113
Difference	-127	-49	431	-255	0	-1	-125	125	0	-3,386	-3,387
Emigration											
Demography	-671	-206	-2,297	-2,429	-15,490	-48,609	-5,684	-2,493	-19,718	-17,834	-115,431
RRC	-2,227	-455	-7334	-3,889	-55,766	-168,556	-10,871	-7,133	-33,689	-31,739	-321,659
Difference	1,556	249	5,037	1,460	40,276	119,947	5,187	4,640	13,971	13,905	206,228
Interprovincial Migration											
Demography	-23,074	1,643	-5,288	-3,255	-51,176	-40,850	-25,336	-26,644	7,155	167,809	984
RRC	-32,767	-886	-1,479	-2,933	-49,395	-37,505	-29,765	-25,095	-10,321	191,222	1,076
Difference	9,693	2,529	-3,809	-322	-1,781	-3,345	4,429	-1,549	17,476	-23,413	-92
Non-permanent Residents											
Demography	-1,406	164	-950	-455	-23,353	-116,602	-1630	-777	-8,267	554	-152,722
RRC	455	236	-549	-606	-13,445	-86,934	-582	144	-5,057	4,890	-101,448
Difference	-1,861	-72	-401	151	-9,908	-29,668	-1,048	-921	-3,210	-4,336	-51,274
Total											
Demography	-9,263	5,483	24,271	13,097	300,849	766,568	24,981	11,098	186,986	466,611	1,790,681
RRC	-17,751	1,583	16,860	11,276	252,014	655,572	6,288	9,312	159,907	472,343	1,567,404
Difference	8,488	3,900	7,411	1,821	48,835	110,996	18,693	1,786	27,079	-5,731	223,277

nationally. Similarly, Newfoundland, Quebec, Alberta and Manitoba, together explain a large part of this difference.

In providing some indication as to the factors responsible for these differences, Table 4 presents comparisons using equation 11 (detailed equation). Alternative estimates are provided on (i) births, (ii) deaths, (iii) immigration, (iv) emigration, (v) interprovincial migration and (vi) net change in the number of non-permanent residents. The most important problems in the explanation of closure error are obvious in Table 4, with specific reference to emigration. As Canada does not have a complete border registration system, emigration is clearly the weakest of all the components to enter into the population estimate program. Without access to direct information on the number of persons leaving Canada, the RRC, with its exhaustive tracing, record linkage and direct interviewing procedures, is considered an improvement over any other data sources currently available. Although there are known problems in the RRC (for example, the previously mentioned frame overlap), the current evaluation points to an obvious error in the postcensal estimates, *i.e.*, an understatement of population outflow from Canada. Overall, the difference as observed nationally (206,228) explains the bulk of the closure error documented in 1996. Similarly with Ontario, difficulties in the estimation of emigration appear to be fundamental (with a difference of fully 119,947).

Without being decisive, the current decomposition also suggests other problematic components beyond emigration in the explanation of closure error for specific provinces. For example, the results suggest that estimates of interprovincial migration might be somewhat misstated for British Columbia and Newfoundland (after acknowledging the differences observed on these components and corresponding closure errors). Overall, an acceptance of the RRC on these more difficult to estimate migratory flows – would not only explain the largest part of this difference in growth – but also the largest part of 1996 closure error. With the closure error that remains, it is useful to turn to the observed difference on enumerated. In so doing, the emphasis shifts away from potential problems in the postcensal estimates.

4.3. Comparisons between Estimates of Enumerated

While the difference in enumerated observed nationally is much smaller than the difference documented on growth, for about half the provinces, this difference is of comparable if not larger size. In interpreting this fact, it is recognized that the RRC was never designed to target the "enumerated" population. With the priority of documenting the number "missed" in the census, the sampling design of the RRC over represents "difficult to enumerate groups"

(for example, single young adults), while under representing persons easily "enumerated". Overall, the comparison on enumerated bears well for the accuracy of the RRC – with non-significant differences across all provinces/territories. Nevertheless, the differences observed in a few provinces are reason for concern, being very close to statistical significance at the 95% level in Quebec (positive difference), and approaching statistical significance in British Columbia, Alberta and Manitoba (negative differences).

In the evaluation of the 1991 coverage study results, two alternative hypotheses have been raised in explanation of differences observed for the enumerated (Royce 1993). At one extreme, it could be argued that all of the difference (for a specific province) be explained in terms of the representativeness of the RRC sample, which implies sampling error or frame deficiencies of one sort or another. At the other extreme, it could be argued that all of the difference be explained due to a failure in documenting the true ratio of enumerated to other classification outcomes, which seems to imply some sort of misclassification error or no trace adjustment bias. A correction for the former of the two hypotheses has a relatively minor impact on the estimate of missed (*i.e.*, all classification outcomes are accordingly inflated or deflated by the proportional difference on enumerated). A correction for the latter could have potentially quite a pronounced impact, as a failure to estimate the true ratio implies that all the difference be assigned to other categories.

If the latter hypothesis applies, a correction potentially reduces the error of closure in nine out of twelve provinces/territories (*i.e.*, in all provinces under which the error of closure is in the same direction as the difference on enumerated). On the other hand, if the differences are due to problems in sample representativeness, a subsequent correction is expected to have negligible impact, if not slightly inflating closure error across most provinces. In addition, the evaluation is complicated by the difficulty in establishing the comparable census figures. Error can be potentially introduced through various sources, including: the census-based estimate of returning emigrants (${}^{91FR}RE_{96PP}$), too much or too little correction for frame overlap, sampling and non-sampling error in the estimation of undercoverage in 1991 and 1996, sampling and non-sampling error in the estimation of overcoverage, and potential error in the classification by province of the enumerated. In this context, further research appears justified as to the true character of errors in the RRC estimate of enumerated.

5. CONCLUSION

In this paper, we have shown how there is additional information available through Canada's census coverage measurement program that is of considerable value in

population estimation. Beyond the ability to estimate census undercount, it is possible to extend the classification results from these studies in order to obtain an alternative estimate of demographic growth – potentially decomposed by component. Using the most important of the coverage studies (*i.e.*, the 1996 Reverse Record Check), a new method was presented which allows for an independent estimate of demographic growth for the intercensal period. The Reverse Record Check not only provides what are considered highly accurate estimates of census coverage error, avoiding some of the correlation biases that have hindered post-enumeration studies in other countries, but also provides very valuable insight as to the magnitude of selected migratory flows of importance to population estimation.

The key to the Reverse Record Check is that it begins with a representative sample of all persons who could have theoretically been in Canada on census day, with only minor deficiencies due to the high quality of vital statistics and immigration data in Canada. Through exhaustive tracing and interviewing procedures, valuable information is then obtained as to the number and characteristics of persons successfully enumerated, missed, counted more than once, as well as useful information on the numbers leaving the country (whether temporarily or permanently), the numbers dying, living in another province, and so on. With a relatively large sample and considerable expertise and effort directed toward minimizing all forms of error, the resultant classification results can potentially inform the population estimates program. This is particularly true with some of the more difficult to estimate migratory flows.

In planning for the 2001 Census, the goal of minimizing all error in the census coverage measurement program remains a priority. As these studies have been designed with a primary target of estimating the population "missed" rather than other classification outcomes (emigrated, deceased, *etc.*), the new demographic approach presented in the current paper leads to the logical question, as to whether its current design need be reworked somewhat if its current usage is broadened. Of interest in this context is the fact that these coverage studies appear to provide an alternative estimate of growth which rivals that as currently available through the population estimates program, and is likely superior with respect to selected components. Further research about how we might more fully exploit this fact appears justified, in improving the quality of the population estimates program.

ACKNOWLEDGEMENTS

We would like to thank R.G. Carter and P. Dick (both of Statistics Canada) and G. Robinson (U.S. Bureau of the Census) for their comments on an earlier draft of this paper. We also acknowledge the helpful comments and suggestions from the Associate Editor and two referees.

REFERENCES

- BRACKSTONE, G.J., and GOSSELIN, J.F. (1973). *Census Evaluation Program, 1971 RRC: Methodology Report*. Statistics Canada. Ottawa, Ontario.
- BURGESS, R.D. (1988). Evaluation of Reverse Record Check estimates of undercoverage in the Canadian census of population. *Survey Methodology*, 14, 137-156.
- CARTER, R.G. (1990). The Measurement of net coverage error in Canadian censuses. *Proceedings: Symposium 90, Measurement and Improvement of Data Quality*, Statistics Canada.
- FELLEGI, I.P. (1969). A theory of record linkage. *Journal of the American Statistical Association*, 64(328), 1183-1210.
- GOSSELIN, J.-F. (1976). The methodology of the 1971 Reverse Record Check. *Survey Methodology*, 2, 180-193.
- ROMANIUC, A. (1988). A demographic approach to the evaluation of undercoverage in the Canadian Census of Population. *Survey Methodology*, 14, 157-172.
- ROYCE, D. (1993). Evaluation of the May 1993 Revised Results of the 1991 Census Coverage Studies. Social Survey Methods Division, Working Paper, Statistics Canada, Ottawa, Ontario.
- ROYCE, D., GERMAIN, M.-F., JULIEN, C., DICK, P., SWITZER, K., and ALLARD, B. (1994). *Coverage: 1991 Census Technical Report*. Catalogue no. 92-341E. Ottawa: Statistics Canada.
- STATISTICS CANADA (1999). *Coverage: 1996 Census Technical Reports*. Catalogue no. 92-370-XPB.
- STATISTICS CANADA (2000). *Annual Demographic Statistics*. Catalogue no. 21-213-XPB.
- TOURIGNY, J., CLARK C., and PROVOST, M. (1998). Evaluation of the March 1998 Preliminary Results of the 1996 Census Coverage Studies. Social Survey Methods Division, Working Paper, Statistics Canada, Ottawa, Ontario.
- TOURIGNY, J., BUREAU, M., and CLARK, C. (1998). Revised Direct Estimates of 1991 Census Coverage Studies. Sept 24th release. Social Survey Methods Division, Working Paper, Statistics Canada, Ottawa, Ontario.

Multilevel Modelling of Complex Survey Longitudinal Data With Time Varying Random Effects

MOSHE FEDER, GAD NATHAN and DANNY PFEFFERMANN¹

ABSTRACT

Longitudinal observations consist of repeated measurements on the same units over a number of occasions, with fixed or varying time spells between the occasions. Each vector observation can be viewed therefore as a time series, usually of short length. Analyzing the measurements for all the units permits the fitting of low-order time series models, despite the short lengths of the individual series. We illustrate this paradigm using simulated data that follow the rotation scheme of the Israel Labor Force Survey (LFS). This survey employs a rotating panel sampling scheme of two quarters in the sample, two quarters out of the sample and then two quarters in again. The model consists of two-level linear models for single time points that are connected by allowing the second level effects (corresponding to households) and the first level residuals (corresponding to individuals) to evolve stochastically over time. The likelihood of the model is easily constructed by employing the time series properties of the combined model. However, in view of the large number of unknown parameters, direct maximization of the likelihood could yield unstable estimators. Therefore, a two-stage procedure is adopted. At the first stage, a separate two-level model is fitted for each time point, thus yielding estimators for the fixed effects and the variances. At the second stage, the time series likelihood is maximized only with respect to the time series model parameters. This two-stage procedure has the further advantage of permitting appropriate first and second level weighting to account for possible informative sampling effects. Empirical results when fitting the model to data collected by the Israel LFS are also presented

KEY WORDS: Informative sampling; Probability weighted IGLS; Rotating panel schemes; State-space models.

1. INTRODUCTION

1.1 Background and Objectives

In recent years there has been a growing interest in fitting models to data collected from longitudinal surveys that use complex sampling designs. This interest reflects expansion in requirements by policy makers and social scientists for in-depth studies of social processes over time, rather than of one-time "snap-shots" provided by cross-sectional analyses. A familiar example is the estimation of gross flows between social and demographic states such as employment states or health and education levels. For discussions of these issues and the problems they raise with respect to the design and analysis of longitudinal surveys, see Duncan and Kalton (1987) and Binder (1998).

Examples of surveys we wish to consider in this paper are of three types:

1. Rotating panel surveys such as labor force surveys carried out in many countries. These surveys were often designed originally for cross-sectional analysis of household and individual data, so as to study labor force and other socio-economic characteristics on a current basis. Complex rotating sampling schemes have later been introduced in order to improve comparisons over time. For example, the quarterly Israel Labor Force Survey (LFS) employs a rotating panel sampling scheme whereby each unit in the

sample is interviewed for two consecutive quarters; it is left out of the sample for the next two quarters and then is interviewed again for two more consecutive quarters. In The U.S.A. and Brazil, a more complicated sampling scheme of 4 months in the sample, 8 months out of the sample and then 4 months in again is used. Australia, Canada and the U.K. employ sampling schemes by which sampled units are interviewed over a succession of months or quarters before being dropped from the sample. These kinds of surveys are increasingly used for short-term longitudinal analysis, such as the estimation of gross flows between labor force states or studies of social mobility. This has not always proved simple due to the complexity of the survey designs, difficulties in matching and response errors.

2. Medium term panel surveys, such as the U.S. Survey of Income and Programme Participation (SIPP, Herriot and Kasprzyk 1984), the U.S. Panel Study of Income Dynamics (PSID, Survey Research Center, 1984) and the Canadian Survey of Labor and Income Dynamics (SLID, Webber 1994). These surveys differ from labor force surveys in being specially designed for longitudinal analysis of economic and social characteristics of households and individuals. For example, SIPP includes an intensive investigation in the form of a full retrospective interview every 4 months. It provides a complete work history for the

¹ Moshe Feder, Department of Social Statistics, University of Southampton, Southampton, S017 1BJ, U.K.; Gad Nathan and Danny Pfeffermann, Department of Statistics, Hebrew University, Jerusalem, 91905, Israel.

survey period (30-48 months) by combining the continuous retrospective four-month recall data with a reconciliation of data provided for longer periods.

3. Longitudinal cohort studies characterized by the follow-up of a cohort sample over a long time period. For example, in the British Household Panel Survey, starting from a sample of addresses selected in 1991, data have been collected on the same households in subsequent annual waves for over seven years. A wide range of data is collected on labor force characteristics, economic resources and health and education, with emphasis on longitudinal aspects. In this survey all members of the originally selected households were followed and the sample was supplemented by the addition of entrants to the sample households, including children born to sample household members. Other longitudinal cohort studies such as the British National Child Development Study and the British Cohort Study have surveyed a cohort of births over periods of up to 40 years. See Nathan (1999) for description and discussion of the latter three studies.

Most of the studies associated with these surveys require longitudinal analysis for populations that have a complex hierarchical structure, based on data collected from complex sampling designs. Standard analysis of longitudinal survey data often fails to account for the complex nature of the sampling design such as the use of unequal selection probabilities, clustering, post-stratification and other kinds of weighting used for the treatment of non-response. The effect of sampling on the analysis is due to the fact that the models in use typically do not incorporate all the design variables determining the sample selection, either because there may be too many of them or because they are not of substantive interest. However, if the design is "informative" in the sense that the outcome variable is correlated with the design variables not included in the model, even after conditioning on the model covariates, standard estimates of the model parameters can be severely biased, leading possibly to false inference. Pfeffermann (1993, 1996) reviews many examples reported in the literature that illustrate the effects of ignoring the sampling process when fitting models to survey data and discusses methods that have been proposed to deal with this problem. See also the book edited by Skinner, Holt, and Smith (1989) and the more recent paper by Pfeffermann, Skinner, Goldstein, Holmes, and Rasbash (1998) to which we refer in more detail below. It should be emphasized that standard inference may be biased even when the original sample design is simple random within design strata, due to non-response, attrition, and imperfect frames that result in *de facto* a posteriori differential inclusion probabilities. Special features of longitudinal studies, such as late additions of individuals who join panel households, can also lead to *de facto* unequal inclusion probabilities.

In this paper we propose to deal with the problems arising from the hierarchical nature of the target population, the longitudinal aspect of the analysis and the effects of complex sampling designs by combining three separate statistical methodologies. These are multilevel modelling (MLM), time series modelling and methods of analysis under complex informative sampling. Multilevel models are used to deal with the hierarchical structure of many human populations like persons within households, pupils within classes, classes within schools and so forth. The models, extensively employed by social scientists especially in the field of education, account for the effects of observed covariates at the lower and higher levels of the structure, with fixed or random coefficients. Common unobservable random effects within the higher levels capture further unexplained variations. The method of Iterative Generalized Least Squares (IGLS) is commonly used for estimating the model parameters, Goldstein (1986, 1995).

Simple state-space time series models are used to combine the multilevel models operating at different time points via a set of linear transition equations that account for the time series relationships of the random covariate coefficients and the higher level random effects. The Kalman filter is used for estimating the model parameters and predict the random effects for current and future time points. Smoothing algorithms can be used for updating past predictions, Harvey (1989). Methods of model fitting under informative sampling are employed to control the bias resulting from the sample selection process. Such methods have been investigated in recent years in the context of analytic inference from complex sample surveys, mostly for cross-sectional analysis of single-level models, *cf.* Skinner *et al.* (1989). In the present paper we utilize the methodology of sample weighting for multilevel modelling as developed by Pfeffermann *et al.* (1998).

The aims of the present study are then to develop models and methods of estimation for longitudinal analysis of hierarchically structured data, taking unequal sample selection probabilities into account. The main feature of our approach is that the model is fitted at the individual level but it contains common higher level random effects that change stochastically over time. The model enables to predict the higher and lower level random effects (like household and individual person effects in the present application), using the data for all the time points with observations. This should enhance model-based inference from complex survey data since it permits a better understanding of the structure and correlation pattern of the longitudinal measurements. In particular, it is bound to improve the prediction of individual measurements compared to the use of aggregate time series models, which by their nature fail to separate the individual (person) effects from the common higher level (household) effects. These advantages are partly illustrated in the example of section 6 and more so in a related paper by Pfeffermann and Nathan (forthcoming) which focuses on the imputation of missing data. It is

important to emphasize in this regard that although the length of each individual longitudinal record is often very short (4 measurements for each individual in our application), the number of records is usually sufficiently large to warrant the application of classical time series estimation and model diagnostic procedures. In this article we only consider parameter estimation under a given model but the use of test statistics and diagnostic procedures that employ the empirical innovations for model identification follows through with minor modifications by virtue of the use of maximum likelihood estimation methods and the consistency of the parameter estimators.

In section 2 we overview the main features of the aforementioned statistical methodologies that are employed in subsequent sections. In section 3 we propose a model that addresses the longitudinal aspects discussed above. Estimation procedures are discussed in section 4. Section 5 contains the results of a simulation study carried out for assessing the performance of the various estimators under different sampling scenarios. Results obtained when fitting the model to real data collected by the Israel LFS are presented in section 6, followed by a brief summary in section 7 of possible model extensions and applications.

1.2 Literature Review

Previous work in this area deals mostly with longitudinal data in a non-survey context and does not consider hierarchically structured populations. In particular, none of the studies that we have come across permits the second level effects (common household effects in our application) to evolve over time. For example, Goldstein, Healy and Rasbash (1994) consider the analysis of repeated measurements using a two-level model with individuals as second levels and the repeated measurements as the first levels. The model extends the standard two-level model by permitting the first level measurements to be correlated over time. The authors consider several possibilities of modelling the autocorrelation structure, which include autoregressive models when the measurements are taken at equally spaced time points and autocorrelation functions when the observations are taken at unequal time intervals. In the latter case the autocorrelation function is linearized for estimation purposes.

Several authors study the application of time series models for the analysis of longitudinal data. In a series of papers by Jones and his co-authors (Jones and Ackerson 1990, Jones and Boadi-Boating 1991, Jones and Vecchia 1993) and the book by Jones (1993), the authors consider observations taken at unequally spaced time gaps. The observations referring to the same subject are allowed to be serially correlated by postulating continuous autoregressive moving average models. These models contain fixed and random effects, but do not have a hierarchical population structure. Weighted least squares and state space modelling combined with the Kalman filter are used for calculating the likelihood function.

Continuous time autoregressive models for irregularly spaced longitudinal data are considered also by Belcher, Hampton and Tunnicliffe (1984), using linear stochastic differential equations for describing the process generating the data. An Empirical Bayes approach is proposed by Bryant and Day (1991) for the simultaneous analysis of a system of mixed linear models, having linked and serially correlated random effects. Chi and Reinsel (1989) consider a score test for autocorrelation between individual errors under a "conditional independence" random effects model. The authors derive a maximum likelihood estimation procedure and use the estimators for predicting the random effects by application of Empirical Bayes.

Diggle, Liang and Zeger (1994) propose the use of generalized linear models for the analysis of longitudinal data. They consider a transition (Markov) model by considering past values as additional predictor variables. Transitional extensions of the GLM are used for maximum likelihood estimation under linear link functions, whereas for non-linear link functions the estimation is based on conditional score functions. Lawless (1999) uses an event history approach for the analysis of longitudinal data. By this approach, the dependent variable is the number of occurrences of a particular event up to a given time point t , with the limiting transitional probabilities being modelled as functions of the previous history and covariates. Zimmerman and Nunez-Anton (1997) propose a structured antedependence model for longitudinal data, primarily in the context of growth analysis. Neither of the above studies considers a hierarchical structure or a complex sampling design.

Finally, Skinner and Holmes (1999) consider a model for longitudinal observations that consists of a "permanent" random effect at the individual level and autocorrelated transitory random effects corresponding to different waves of investigation. The authors study two approaches for the estimation of the unknown model parameters with both approaches accounting for sampling effects and "non informative" attritions. The first approach treats the repeated observations as correlated multivariate outcomes and derives probability-weighted estimators that account for the correlation structure. The second approach considers the model as a two-level model with "individuals" as the second level units and the repeated measurements as first level units. Estimation of the unknown parameters under this approach is carried out by a modification of the PWIGLS method of Pfeiffermann *et al.* (1998, see section 2.2).

2. STATISTICAL METHODOLOGIES UNDERLYING THE PROPOSED APPROACH

2.1 Multilevel Models

In what follows we consider a two-level model for the response variable y in a population consisting of

$i = 1, \dots, M$ second level units (household, schools, ...) and $j = 1, \dots, N_i$ individuals within second level unit i . The model is,

$$y_{ij} = x'_{ij}\beta + z'_{ij}u_i + z_{0ij}e_{ij}, \quad i = 1, \dots, M; j = 1, \dots, N_i, \quad (2.1)$$

where x_{ij} , z_{ij} , and z_{0ij} are known covariate values of dimensions p , q and 1 respectively, β is a fixed parameter vector of dimension p and $u_i \sim N(0, \Omega)$ and $e_{ij} \sim N(0, \sigma^2)$ are independent random second level effects and first level residuals of orders p and 1 respectively.

The inclusion of the multipliers z_{0ij} allows for first level heteroscedasticity whereas the common second level effects u_i explain the (interclass) correlations between individual measurements corresponding to the same second level unit. In the simple case of the "random intercept model", $y_{ij} = x'_{ij}\beta + u_i + e_{ij}$, these correlations take the familiar form, $\text{Corr}(y_{ij}, y_{ik}) = \sigma_u^2 / (\sigma_u^2 + \sigma^2)$. The random intercept model is often applied for small area estimation (see below).

As stated in the introduction, models like (2.1) are widely used by social scientists for studying the effects of the covariate variables and the interrelationships between observations corresponding to the same higher level unit. In such cases, primary interest is in the estimation of the vector coefficient β and the vector θ of the distinct elements of Ω and σ^2 . Another, well-known application of the two-level model is for "small area estimation", in which case the second levels are geographical areas or other domains of study. In small area estimation, the target of the analysis is the prediction of the second level (area) means $\bar{X}'_i\beta + \bar{Z}'_i u_i$, where \bar{X}_i and \bar{Z}_i are the true area covariate means, and the estimation of the model parameters is only an intermediate step. See Rao (1999) for a recent review.

Estimation of the unknown model parameters is carried out most conveniently by use of the Iterative Generalized Least Squares (IGLS) algorithm (Goldstein 1986, 1995). For a random sample of m second level units and n_i first level units within second level unit i , the model holding for the sample data is first written in matrix form as

$$y_i = X_i\beta + d_i, \quad i = 1 \dots m \quad (2.2)$$

where $y_i = [y_{i1}, \dots, y_{in_i}]'$, $X_i = [x_{i1}, \dots, x_{in_i}]'$ and $d_i = [d_{i1}, \dots, d_{in_i}]'$ with $d_{ij} = (z'_{ij}u_i + z_{0ij}e_{ij})$. Then, $d_i \sim N(0, V_i)$, where $V_i = Z_i\Omega Z_i' + \sigma^2 Z_{0i}Z_{0i}' = V_i(\theta)$; $Z_i = [z_{i1} \dots z_{in_i}]'$ and $Z_{0i} = \text{diag}[z_{0i1} \dots z_{0in_i}]$. The IGLS algorithm iterates between the estimation of β , with θ considered known, and the estimation of θ , with β considered known. At each iteration, the estimate obtained for the other vector parameter on the previous iteration is used as the "known" parameter. This process is a special case of the EM algorithm and it converges to the corresponding maximum likelihood estimators (MLE) under the stated normality assumptions. It is known to provide consistent estimators under more general conditions.

2.2 MLM Estimation Under Informative Sampling

The IGLS algorithm described in section 2.1 assumes that the model defined by (2.2) holds for the sample data. This would be the case if selection of the first and second level units is carried out by simple random sampling. However, as discussed in the introduction, the selection of the sample could be informative so that the model holding for the sample units differs from the model holding in the population. For example, in an educational survey, schools in poor areas could be sampled with higher probabilities. In a household survey, higher selection probabilities could be assigned to households in areas characterized by high proportions of minorities or to persons that are unemployed. As illustrated by Pfeffermann *et al.* (1998) and also in section 5 of the present paper, the use of the IGLS algorithm in such cases could yield severely biased estimators for all the parameters. The authors propose therefore a probability weighted IGLS (PWIGLS) algorithm that protects against informative sampling.

The algorithm is an adaptation of the pseudo-MLE method (Binder 1983, Skinner *et al.* 1989, Pfeffermann 1993). Suppose that the two-level model defined by (2.1) holds for the target population. Had all the population values been observed, the IGLS would converge at the end of the iterative process to the census estimators, $(\hat{\beta}_c, \hat{\theta}_c)$. At each iteration, the intermediate estimators $(\hat{\beta}_{(i)}, \hat{\theta}_{(i)})$ are products of matrices with elements that are functions of sums of the population values. When the IGLS is applied to sample data, the population sums are substituted by the corresponding sample sums. The PWIGLS consists of further replacing the unweighted sample sums by weighted sums. Denote by $\pi_i = \Pr(i \in s)$ the second level sample inclusion probabilities and by $\pi_{j|i} = \Pr(j \in s | i \in s)$ the conditional first level inclusion probabilities. The PWIGLS estimators are obtained by, 1- replacing each second level sample sum of the general form $\sum_{i=1}^n g_i$ by the weighted sum $\sum_{i=1}^n w_i g_i$, where $w_i = \pi_i^{-1}$ and 2- replacing each first level sample sum $\sum_{j=1}^{n_i} g_{ij}$ by the weighted sum $\sum_{j=1}^{n_i} w_{j|i} g_{ij}$ with $w_{j|i} = \pi_{j|i}^{-1}$. Note that the weighting process requires the knowledge of the inclusion probabilities at both stages of the selection process and not just the final overall inclusion probabilities $\pi_{ij} = \pi_{j|i} \times \pi_i$.

As established by Pfeffermann *et al.* (1998), the PWIGLS estimators are consistent for the model parameters when both the first and second level sample sizes increase, but the estimators of the variances are not consistent if the first level sample sizes are bounded. For this case, the authors propose appropriate scaling of the weights $w_{j|i}$ that eliminates the bias, provided that the sample selection within the second level units is noninformative. It is important to emphasize that standard weighting of the sample measurements by the weights $w_{j|i} = \pi_{j|i}^{-1}$, which is routinely applied for single level models yields consistent estimators only for β .

2.3 State-space Models

State-space models as considered here consist of two sets of equations:

1. The measurement (observations) equation:

$$y_t = X_t \beta_t + L_t \alpha_t + \varepsilon_t; E(\varepsilon_t) = 0, \\ E(\varepsilon_t \varepsilon_{t-k}') = \delta_k H_t, t = 1, \dots, T \quad (2.3)$$

2. The transition (system) equation:

$$\alpha_t = G_t \alpha_{t-1} + \eta_t; E(\eta_t) = 0, \\ E(\eta_t \eta_{t-k}') = \delta_k Q_t, t = 1, \dots, T \quad (2.4)$$

where $\delta_k = 1$ for $k = 0$ and $\delta_k = 0$ otherwise. We also assume $E(\varepsilon_t \varepsilon_s') = 0$ for all t and s . Note that both y_t and α_t can be multivariate. The measurement equations relate the observations y_t at any given time point to covariate values X_t with fixed (nonstochastic) vector coefficients β_t , and linear functions L_t of an unobservable state vector α_t . The transition equations describe the time series relationships between the components of the state vector. The matrices X_t , L_t and G_t are assumed to be nonstochastic although they may change over time, as is the case with the vector coefficients β_t . Notice that the latter vectors can be included as part of the state vectors by taking their transition matrix to be the zero matrix of corresponding order and defining the corresponding residual variances in Q_t to be very large. See Sallas and Harville (1981) for details.

Although not written here in its most general form, the state-space model defined by (2.3) and (2.4) is known to include as special cases many of the time series and mixed linear models in common use. As important examples we mention the family of ARIMA models and models with random regression coefficients. The MLM defined by (2.1) can also be easily structured in a state-space form. To see this, replace the index i by t and define $L_t = [X_t, Z_t]$, $\alpha_t = [\beta_t', u_t']'$, $H_t = \sigma^2 Z_{0t}$ and $G_t = [I_p, 0_q]$ where I_p and 0_q define the identity matrix and the zero matrix of the appropriate orders. (The matrices Z_t and X_t are defined below (2.2).) The vector coefficient β_t is added for convenience to the state vector. The covariance matrix Q_t is block diagonal with 0_p and $Z_t \Omega Z_t'$ as the two blocks. The use of the zeroes matrix 0_p for the covariance of $(\beta_t - \beta_{t-1})$ guarantees that the β -coefficients are fixed over time, in accordance with (2.1). (The representation of the MLM in a state-space form is not unique.)

For given covariance matrices $\{H_t, Q_t\}$ and assuming that β_t, L_t and G_t are known for all t , the best linear unbiased predictor (BLUP) of the state vector at any given time t , based on all the data accumulated until that time, is conveniently obtained by means of the Kalman Filter. Let $\hat{\alpha}_{t-1}$ define the BLUP of α_{t-1} based on the observations until time $(t-1)$, with covariance matrix $P_{t-1} =$

$\text{Cov}(\hat{\alpha}_{t-1} - \alpha_{t-1})$. The BLUP of α_t at time $(t-1)$ is then, $\hat{\alpha}_{t|t-1} = G_t \hat{\alpha}_{t-1}$ with covariance matrix $P_{t|t-1} = \text{Cov}(\hat{\alpha}_{t|t-1} - \alpha_t) = G_t P_{t-1} G_t' + Q_t$. When new observations y_t become available, the predictor $\hat{\alpha}_{t|t-1}$ and the corresponding covariance matrix are updated as

$$\hat{\alpha}_t = \hat{\alpha}_{t|t-1} + P_{t|t-1} L_t' F_t^{-1} (y_t - X_t \beta_t - L_t \hat{\alpha}_{t|t-1}) \\ P_t = P_{t|t-1} - P_{t|t-1} L_t' F_t^{-1} L_t P_{t|t-1} \quad (2.5)$$

where $F_t = L_t P_{t|t-1} L_t' + H_t = \text{Var}(y_t - \hat{y}_{t|t-1})$ with $\hat{y}_{t|t-1} = X_t \beta_t + L_t \hat{\alpha}_{t|t-1}$ defining the BLUP of y_t at time $(t-1)$. The actual application of the Kalman filter requires a proper initialization for $\hat{\alpha}_{1|0}$ and $P_{1|0}$ which depends on the model under study. See section 4 for the initialization under the model proposed in this paper.

The unknown model parameters (β_t , elements of H_t, Q_t and possibly L_t and G_t) are ordinarily estimated by MLE with the likelihood conveniently constructed by use of the "prediction error decomposition". Assuming that $\dim(y_t) = n$, the log-likelihood takes the general form,

$$\log(L) = -\{T \frac{n}{2} \log(2\pi) + \frac{1}{2} \sum_{t=1}^T \log |F_t| \\ + \frac{1}{2} (Y_t - \hat{Y}_{t|t-1})' F_t^{-1} (Y_t - \hat{Y}_{t|t-1})\}. \quad (2.6)$$

For a thorough discussion of state-space models and their applications, see Harvey (1989).

3. A MODEL FOR HIERARCHICAL LONGITUDINAL DATA

In this section we propose a time series multilevel model which combines separate cross-sectional two-level models by modelling the evolution of the first and second level random effects over time. Let S_t define the sample available at time t , composed of m_t level 2 units with n_h level 1 units in level 2 unit h . The formulation of the overall sample in terms of the subsets S_t covers situations where the longitudinal observations are collected at different time periods. The proposed model allows also for the rotation patterns mentioned previously and for wave non-response. Note that the samples observed at different time points are generally not disjoint and that the assumption that n_h is fixed over time is not restrictive. Pfeffermann and Nathan (forthcoming) consider the case of temporal missing data for which this supposition does not hold. As long as the missing data are missing completely at random, generalization of the present methodology to this case is straightforward. We assume the following two-level model to hold for the sample S_t :

$$y_{hjt} = x'_{hjt} \gamma_t + z'_{ht} v_t + z'_{ht} u_{ht} + e_{hjt}, \\ h = 1, \dots, m_t, j = 1, \dots, n_h, \quad (3.1)$$

where y_{hjt} is the outcome for first level unit j in second level unit h , x_{hjt} and z_{ht} are fixed known covariate vectors of dimensions p and q respectively, γ_t and v_t are fixed (unknown) vector coefficients and u_{ht} and e_{hjt} are independent second level and first level random effects. For given time t , The model defined by (3.1) is basically the same as the MLM model defined by (2.1), except that we assume $z_{hjt} = z_{ht}$ for all j and t , thus distinguishing between first level covariates and second level covariates. We assume also for convenience $z_{0hjt} = 1$. The model is quite general in that all the covariate variables, the fixed vector coefficients and the random effects are allowed to vary over time in ways defined below. Notice that by assuming that (3.1) holds for the sample data, it is implicitly assumed that the sampling design is noninformative. See the discussion in section 2.2 and also section 4 below.

As in (2.2), the model defined by (3.1) can be formulated in matrix form as,

$$Y_{ht} = X_{ht} \gamma_t + Z_{ht} v_t + Z_{ht} \mu_{ht} + I_{n_h} e_{ht}, \quad (3.2)$$

where $Y_{ht} = [y_{h1t}, \dots, y_{hn_h t}]'$, $X_{ht} = [x_{h1t}, \dots, x_{hn_h t}]'$, $Z_{ht} = 1 \otimes z_{ht}$ and $e_{ht} = [e_{h1t}, \dots, e_{hn_h t}]'$ with \otimes defining the Kronecker product. The matrix representation (3.2) can be written concisely as,

$$Y_{ht} = \tilde{X}_{ht} \beta_t + \tilde{Z}_{ht} \alpha_{ht}, \quad (3.3)$$

where $\tilde{X}_{ht} = [X_{ht}, Z_{ht}]$; $\tilde{Z}_{ht} = [Z_{ht}, I_{n_h}]$; $\beta_t = [\gamma_t', v_t']'$; $\alpha_{ht} = [u_{ht}', e_{ht}']'$.

Next we model the time series relationships of the vector coefficients and the random effects. We assume that the vectors β_t , $t = 1, 2, \dots$ are fixed without specifying the way they evolve over time. This assumption is generally not restrictive because in practical applications the overall sample size in any given time point is usually sufficiently large to allow accurate estimation of the vector coefficients without having to borrow information across time. For the random second and first level effects we postulate first order autoregressive [AR(1)] relationships of the form,

$$u_{ht} = A u_{h,t-1} + \delta_{ht}; \quad e_{ht} = \rho e_{h,t-1} + \varepsilon_{ht} \quad (3.4)$$

where A is a $(q \times q)$ matrix of fixed coefficients, ρ is a fixed scalar and $\delta_{ht} \sim N(0, \Delta)$; $\varepsilon_{ht} \sim N(0, \sigma_e^2 I_{n_h})$ are independent white noise series. The model defined by (3.4) is rather simple and as a further simplification we assume that A and Δ are diagonal, implying that the second level random effects are independent. It is assumed also that $|\rho| < 1$ and $|A_{kk}| < 1$ for all k to guarantee stationarity. More complex models can be considered in principle but it should be emphasized that unlike in classical (aggregate) time series analysis, longitudinal observations may only be taken over a very short time period in which case the use of models that incorporate lagged values of high order may no longer be operational. For example, in the quarterly Israel

LFS described in the introduction, individuals are in the sample for a total of 4 quarters over a time period of 6 quarters which clearly limits the class of time series models that can be postulated for the random effects.

The AR(1) models defined by (3.4) can be written concisely as

$$\alpha_{ht} = G_h \alpha_{h,t-1} + \eta_{ht}, \quad h = 1, \dots, m_t \quad (3.5)$$

where,

$$G_h = \begin{bmatrix} A & 0 \\ 0 & \rho I_{n_h} \end{bmatrix}, \quad \eta_{ht} = \begin{bmatrix} \delta_{ht} \\ \varepsilon_{ht} \end{bmatrix},$$

$$\eta_{ht} \sim N(0, Q_h), \quad Q_h = \begin{bmatrix} \Delta & 0 \\ 0 & \sigma_e^2 I_{n_h} \end{bmatrix}. \quad (3.6)$$

By writing the proposed model using the equations (3.3), (3.5) and (3.6) and setting $\tilde{Z}_{ht} = L_{ht}$, $H_{ht} = 0$, it is easily seen to belong to the class of state-space models presented in section 2.3, with no residual errors in the measurement equation. The model is defined for distinct second level units h but unlike in classical time series analysis where the data consist of a single long series, the data in our case consist of many independent short (longitudinal) series that could be observed over different time periods. Note that the transition matrix, G_h and the covariance matrix, Q_h depend on h through the second level size n_h but they are time invariant. In situations where the second level sizes are not fixed over time (for example, because of missing data), these matrices also change accordingly.

4. ESTIMATION OF THE MODEL PARAMETERS

In principle, the likelihood function holding for the model defined by (3.3), (3.5) and (3.6) can be maximized to obtain the maximum likelihood estimators (MLE) of all the unknown model parameters. However, the number of estimated parameters would usually be very large, which can intensify the computations and result in statistically unstable estimators. For instance, even for $p = q = 2$ and $T = 10$ there are already 46 unknown parameters. We propose therefore a two-stage estimation procedure that employs MLM estimation for the "cross-sectional parameters" and state-space model estimation for the "time series parameters". The use of this procedure has the further advantage of accommodating appropriate weighting to protect against informative sampling.

The procedure starts off by fitting the MLM defined by (3.1) to each sample S_t separately, to obtain IGLS estimates of the time-dependent fixed effects $\beta_t = [\gamma_t', v_t']'$ and the variances of the random effects u_{ht} and e_{hjt} . Notice that by (3.4),

$$\text{Var}(u_{ht}) = \Delta^* = (I - A^2)^{-1} \Delta;$$

$$\text{Var}(e_{hjt}) = \sigma_e^2 = (1 - \rho^2) \sigma_e^2 \quad (4.1)$$

using familiar relationships holding for AR(1) models. The use of this step yields estimates $\{\hat{\beta}_t, \hat{\Delta}_t^*, \hat{\sigma}_{et}^2\}$ for $\{\beta_t, \Delta^*, \sigma_e^2\}$ respectively. Under the model, the true variances (Δ^*, σ_e^2) are fixed over time and assuming that the sample sizes at the various time points are fairly constant, the estimates $\hat{\Delta}_t^*$ and $\hat{\sigma}_{et}^2$ can be averaged to yield single estimates

$$\bar{\Delta}^* = \sum_{t=1}^T \hat{\Delta}_t^* / T; \quad \bar{\sigma}_e^2 = \sum_{t=1}^T \hat{\sigma}_{et}^2 / T. \quad (4.2)$$

In the second stage the remaining parameters are estimated by maximizing the likelihood of the combined model defined by (3.3) (3.5) and (3.6), with the parameters estimated in the first stage held fixed at their estimated values. Since observations on different second level units are independent, the log-likelihood has the form $\log(L) = \sum_h \log(L_h)$ where L_h , the contribution to the likelihood from second level unit h , is defined by (2.6) with the index h added to all the components thus distinguishing between different second level units. As pointed out before, the number of time points for which the second level units are observed and the time periods over which the observations are taken may differ between units so that the notation T in (2.6) for the number of time points needs also to be changed to T_h .

When fitting the model to data obtained from rotating panel sampling designs as in the empirical study of the present paper, a further modification is required to account for the intermediate periods without observations. For example, for the Israel LFS described in the introduction, with rotation pattern of two quarters in the sample, two quarters out of the sample and two quarters in again, $T_h = 4$ but the transition equations from $t=2$ to $t=3$ (the next quarter with observations) have to be changed to account for the two quarters with missing observations. Repeated substitutions in (3.5) yield the following relationships:

$$\alpha_{h3} = G_h^3 \alpha_{h2} + \eta_{h3}^*; \quad \eta_{h3}^* \sim N(0, Q_{h3}^*),$$

$$Q_{h3}^* = \begin{bmatrix} (A^4 + A^2 + I)\Delta & 0 \\ 0 & (\rho^4 + \rho^2 + 1)\sigma_e^2 I_{n_h} \end{bmatrix}. \quad (4.3)$$

In order to apply the Kalman filter and compute the likelihood, it is needed to set initial values for α_{10} and P_{10} . This is simple under the present model as $\alpha_{ht} = [u_{ht}', e_{ht}']'$ is stationary with zero mean and covariance matrix defined by (4.1). Thus, the filter is started by setting,

$$\alpha_{h10} = E(u_{h1}', e_{h1}') = 0;$$

$$P_{h10} = \text{Var}[u_{h1}', e_{h1}']$$

$$= \text{diag}\{(I - A^2)^{-1} \Delta, \sigma_e^2 (1 - \rho^2)^{-1} I_{n_h}\}. \quad (4.4)$$

In the empirical study described in the next two sections we compare two methods regarding the set of parameters estimated in the second stage.

Method 1: The parameters estimated in Stage 2 are the three AR coefficients ρ, A_{11}, A_{22} and the corresponding residual variances $\sigma_e^2 = \text{Var}(e_{hjt})$ and $\Delta = \text{Var}(\delta_{ht})$, (equation 3.6, three variances in total). Note that under this method the only estimates utilized from Stage 1 are the fixed parameter estimates $\{\hat{\beta}_t = [\hat{\gamma}_t', \hat{\nu}_t']'\}$. By (4.1), the variances $\Delta^* = \text{Var}(u_{ht})$ and $\sigma_e^2 = \text{Var}(e_{hjt})$ are estimated as

$$\hat{\Delta}^* = (1 - \hat{A}^2)^{-1} \hat{\Delta}; \quad \hat{\sigma}_e^2 = (1 - \hat{\rho}^2)^{-1} \hat{\sigma}_e^2. \quad (4.5)$$

Method 2: The only parameters estimated in Stage 2 are the AR coefficients ρ, A_{11}, A_{22} (Equation 3.4). Note that with this method the variances Δ and σ_e^2 are set in the likelihood as, $\Delta = (I - A^2) \bar{\Delta}^*$ and $\sigma_e^2 = (1 - \rho^2) \bar{\sigma}_e^2$ utilizing (4.1), where $\bar{\sigma}_e^2$ and $\bar{\Delta}^*$ are defined by (4.2).

The estimation procedures described so far assume implicitly noninformative sampling. As discussed in the introduction, complex sample surveys often involve selection with unequal probabilities that could be correlated with the values of the response variable. When this is the case, the model holding for the sample data may differ from the model holding in the population. A further advantage of the proposed two-stage estimation method is that it can be adapted to protect against informative sampling. This is done by applying the weighting procedure described in section 2.2 in the first stage, replacing the iterative IGLS algorithm by the PWIGLS procedure. Thus, for each sample S_t , PWIGLS is used for estimating the MLM model parameters instead of using the IGLS.

Comment 1: Informative selection of the first and second level units does not affect the conditional distributions of the random effects as defined by (3.4). Thus, although the distribution of u_{h1} and e_{h1} could be largely distorted because of the sample selection at time $t = 1$, this has no effect on the distributions of $u_{h2}|u_{h1}$, or $e_{h2}|e_{h1}$. The implication of this property is that the computation of the likelihood in the second stage remains the same, but care should be taken of a proper initialization of the Kalman filter. As defined by (4.4), the filter is initialized by the unconditional means and variances of the random effects under the model, but at time $t = 1$ the moments holding for units in the sample can be different because of the sampling effects. As is well known, for long enough series and under

some regularity conditions, the estimates derived from maximization of the likelihood are not sensitive to the initialization procedure but with short series, improper initialization under informative sampling could distort the estimation process. Nonetheless, as illustrated in section 5, having a moderate number of longitudinal observations even of very short length (at most 4 observations in our application) and weighting the likelihood contributions by the inverse of the sample inclusion probabilities (application of the pseudo likelihood approach) yields approximately unbiased estimators for all the time series model parameters.

5. SIMULATION RESULTS

In this section we report the results of a Monte Carlo study carried out for assessing the performance of the various estimation procedures described in section 4 under noninformative and informative rotating sampling schemes.

5.1 Description of Simulation Study

A) Generation of population data and sample rotation scheme

Population values have been generated for individuals (first level units) within households (second level units), using the model defined by (3.1) and (3.4) (see below). The number of persons n_h observed within household h was selected at random with possible values of 2, 3 or 4. A new panel of households has been generated in each of 11 quarters and a sample of these households has been observed following the Israel Labor Force Survey rotation scheme of two quarters in the sample, two quarters out of the sample and two quarters in again. As easily checked, this process yields a complete sample of four panels in each of the quarters 6-11, with one panel in each quarter observed for the first time, one panel observed for the second time, one for the third time and one for the fourth and last time. (In the first quarter there is only one panel, in the next three quarters there are two panels and in the fifth quarter there are 3 panels.) In what follows we only consider the data observed for quarters 6-11.

B) Population model

The model used for generating the y -values for a given household h is defined by (3.1) and (3.4) with $x'_{hit} = (x_{h1}, x_{h2})$ and $z'_{ht} = (z_{h2})$, such that the covariate values are fixed over time. The x -values were generated independently from the uniform distribution $U[1, 2]$. Values z_{h2} were generated from the uniform distribution $U[1, 5]$. In order to simplify the presentation and evaluation of the results, we also set the model coefficients to be time invariant such that $\gamma_t = \gamma = (6, -2)'$ and $v_t = v = (1, 2)'$. The random error

terms were generated independently between households using the model (3.4) with $A = \text{diag}[0.5, 0.7]$, $\Delta = \text{diag}[0.8, 0.5]$, $\rho=0.4$ and $\sigma_e^2 = 0.25$. Notice from (4.1) that $\text{Var}(u_{ht}) = \Delta^* = \text{diag}[1.067, 0.980]$ and $\text{Var}(e_{hit}) = \sigma_e^2 = 0.298$.

C) Sample selection

We consider two separate sampling schemes.

C1) Noninformative sampling:

Population values have been generated for panels of 30 households, with all the households belonging to a given panel selected to the sample and observed following the sample rotation scheme described in A above. The total number of sampled households in each of the quarters 6-11 is therefore $m = 120$. All the individuals belonging to a given household have been observed, yielding an expected sample size of $n = 360$ individuals for each of the quarters. This sampling scheme corresponds to simple random sampling of households and individuals within the selected households.

C2) Informative sampling

Population values have been generated for panels of 55 households. Households with random effects $u_{h1,1} < 0$ (the value of the first random effect at the first time point) have been sampled with probability 1, households with random effects $u_{h1,1} > 0$ have been sampled independently (Poisson sampling) with probability 0.1. All the individuals belonging to a sampled household have been observed. This sampling scheme yields an expected sample size of approximately 30 households per panel and expected sample sizes of approximately $m = 120$ households and $n = 360$ individuals per quarter, similarly to the sampling scheme C1.

Comment 2: It should be emphasized that even though there are 4 panels observed in each of the quarters 6-11, there are only 11 separate panels that are used for estimation of the model parameters. Moreover, out of the 11 panels, only the panel entering the sample in quarter 6 for the first time is observed in 4 quarters, only 2 panels are observed in 3 quarters, 6 panels are observed in 2 quarters and 2 panels are observed in only one quarter. This implies a total of 13 panel transitions, with about 390 household transitions observed for estimation of the time series parameters. (By a panel transition we mean that the same panel is observed on two occasions. For 3 of these panel transitions there is a time gap of 2 quarters between the two observations). We refer to this sample structure when assessing the estimation of the time series model parameters.

The whole process of generating population values and selecting the sample has been repeated 100 times for each of the two sampling schemes C1 and C2, with one sample selected from each population. For each sample we applied

the two estimation procedures described in section 4. The simulations were run using the Gauss software package. Maximization of the likelihood has been carried out using the numerical optimization procedure, OPTMUM.

5.2 Results

The results of the simulation study are summarized in Tables 1-4 as averages over the 100 samples selected under the two sampling schemes. Each table contains the mean estimates of the model parameters, the empirical standard deviations (SD) of the estimators and the conventional *t*-statistics obtained by dividing the difference between the mean estimates and the true parameter values by the standard errors (SE), computed as SD/10. Notice that the estimates of the fixed vector coefficients $\beta_t = (\gamma_t', v_t')'$ are the same under the two estimation methods.

Perhaps the most important outcome of this study, revealed from Table 1, is that under noninformative sampling it is indeed possible to fit successfully simple but nontrivial time series models to very short longitudinal series, provided that the number of observed series is sufficiently large. (The model is not trivial because even after subtracting the fixed effects, the dependent response variable is the sum of three AR(1) processes.) This conclusion is further strengthened by the fact that 8 out of the 11 panels have been observed for at most 2 times, yielding a total of 13 panel transitions, three of which with a gap of 2 quarters. See Comment 2 at the end of section 5.1.

Next we consider the case of informative sampling. Table 2 shows the results obtained when ignoring the informative sampling process, using the same estimation procedures as used for the noninformative case. As indicated very clearly by this table, some of the parameter estimates are highly significant, particularly the estimators of the parameters indexing the time series model of the random effects u_{h1t} that define the sample selection probabilities. Thus, we find that the absolute relative bias in estimating v_1 is about 27%, and large absolute relative biases are also observed for the estimators of A_{11} and Δ_{11}^* . (The model defined by (3.1) can be rewritten as $y_{hjt} = x_{hjt}'\gamma_t + z_{ht}'u_{ht}^* + e_{hjt}$ where $u_{ht}^* = u_{ht} + v_t$, such that for $v_t \equiv v$ as under the simulation model, $v_1 = E(u_{h1t}^*)$). Note that the three biases are negative, which is explained by the fact that the selection mechanism utilized for this study oversamples individuals with observations that contain negative random effects $u_{h1,1}$. In this case again, the two estimation methods perform very similarly.

Table 2
Means, Standard Deviations (SD) and *t*-Statistics of Estimators Under Two Estimation Methods. Informative Sampling, Unweighted Estimators

Parameter	True Value	Method 1			Method 2		
		Mean	SD	<i>t</i> -statistic	Mean	SD	<i>t</i> -statistic
γ_1	6.000	5.998	0.02	-0.768	5.998	0.02	-0.768
γ_2	-2.000	-2.000	0.03	0.104	-2.000	0.03	0.104
v_1	1.000	0.728	0.09	-34.385	0.728	0.09	-34.385
v_2	2.000	2.005	0.09	0.564	2.005	0.09	0.564
A_{11}	0.500	0.438	0.09	-6.742	0.434	0.09	-7.453
A_{22}	0.700	0.738	0.09	4.078	0.735	0.09	3.941
Δ_{11}^*	1.067	0.995	0.09	-7.766	0.994	0.09	-7.883
Δ_{22}^*	0.980	1.003	0.10	2.352	0.987	0.10	0.698
ρ	0.400	0.407	0.02	3.184	0.405	0.02	2.218
σ_e^2	0.298	0.298	0.01	0.644	0.296	0.01	-1.800

Table 3 shows the results obtained when using the PWIGLS algorithm for the estimation of the MLM parameters (section 2.2) and weighting the time series likelihood contributions $\log(L_h) = -\{1/2 T_h n_h \log(2\pi) + 1/2 \sum_{t=1}^{T_h} \log |F_{ht}| + 1/2 (Y_{ht} - \hat{Y}_{h1t-1})' F_{ht}^{-1} (Y_{ht} - \hat{Y}_{h1t-1})\}$ by the household sampling weights $w_h = 1 / \text{Pr}(h \in s)$, using the same 100 samples as used for Table 2. Weighting the likelihood contributions by the inverse of the sample inclusion probabilities is an application of the pseudo likelihood approach that is often recommended for fitting single level models to cross-sectional data, see, e.g., Binder (1983), Skinner *et al.* (1989) and Pfeffermann (1993). As revealed from this table, the use of the PWIGLS algorithm and weighting the likelihood eliminates the large biases observed in Table 2, despite the improper initialization of the Kalman filter with very short series. (See the discussion in Comment 1 at the end of section 4.) Here again, the two estimation methods perform quite similarly, yielding one

Table 1

Means, Standard Deviations (SD) and *t*-Statistics of Estimators Under Two Estimation Methods. Noninformative Sampling

Parameter	True Value	Method 1			Method 2		
		Mean	SD	<i>t</i> -statistic	Mean	SD	<i>t</i> -statistic
γ_1	6.000	6.002	0.03	0.677	6.002	0.03	0.677
γ_2	-2.000	-2.000	0.03	0.078	-2.000	0.03	0.078
v_1	1.000	0.989	0.08	-1.357	0.989	0.08	-1.357
v_2	2.000	2.008	0.08	0.997	2.008	0.08	0.997
A_{11}	0.500	0.497	0.07	-0.391	0.491	0.07	-1.271
A_{22}	0.700	0.696	0.07	-0.532	0.695	0.07	-0.820
Δ_{11}^*	1.067	1.054	0.08	-1.668	1.045	0.08	-2.677
Δ_{22}^*	0.980	0.991	0.10	1.042	0.990	0.11	0.906
ρ	0.400	0.398	0.02	-0.937	0.397	0.02	-1.637
σ_e^2	0.298	0.298	0.01	-0.062	0.297	0.01	-1.382

Evaluation of the performance of the two sets of estimators in Table 1 shows that all the estimators under Method 1 are highly insignificant based on the conventional *t*-statistics and only the estimator of Δ_{11}^* is significant under Method 2. Note that even in that case the absolute relative bias is about 2% and considering that MLE of time series parameters are generally not strictly unbiased, such a small bias in one of 10 parameters is expected. Notice also that the standard errors of the mean estimators under the two methods are very similar, a result observed also in the other tables.

biased estimator in each case but with both biases being relatively very small.

It is important to mention that the SD's of the weighted estimators shown in Table 3 are always larger than the corresponding SD's of the unweighted estimators displayed in Table 2. As pointed out by one of the referees, this implies that the empirical root mean square errors (RMSE's) of the unweighted estimators in Table 2 are in fact larger than the empirical RMSE's of the corresponding estimators in Table 3. This outcome, however, is due to the relatively small sample sizes employed in this study. For larger samples (larger numbers of households and individuals within the households) the RMSE is dominated by the bias which, unlike the variance, is not reduced as the sample size increases. Thus, it is clear that as the sample size increases the RMSE's of the weighted estimators become smaller than the RMSE's of the unweighted estimators. The fact that probability weighted estimators have larger variances than the corresponding unweighted estimators is well known from many other studies, see Pfeffermann (1993) for discussion and references.

Table 3

Means, Standard Deviations (SD) and *t*-Statistics of Estimators Under Two Estimation Methods. Informative Sampling, Weighted Estimators

Parameter	True Value	Method 1			Method 2		
		Mean	SD	<i>t</i> -statistic	Mean	SD	<i>t</i> -statistic
γ_1	6.000	5.997	0.04	-0.607	5.997	0.04	-0.607
γ_2	-2.000	-2.000	0.05	-0.007	-2.000	0.05	-0.007
v_1	1.000	0.978	0.14	-1.518	0.978	0.14	-1.518
v_2	2.000	2.019	0.14	1.330	2.019	0.14	1.330
A_{11}	0.500	0.490	0.15	-0.695	0.477	0.14	-1.611
A_{22}	0.700	0.699	0.17	-0.066	0.709	0.16	0.545
Δ_{11}^*	1.067	1.055	0.17	-0.664	1.040	0.17	-1.560
Δ_{22}^*	0.980	1.023	0.19	2.199	1.010	0.19	1.571
ρ	0.400	0.401	0.04	0.135	0.397	0.04	-0.813
σ_e^2	0.298	0.297	0.01	-0.486	0.294	0.01	-3.340

As discussed in Comment 1 at the end of section 4, informative sampling distorts the cross-sectional distribution of the sample observations and the initialization of the Kalman filter, but does not affect the conditional distributions of the first and second level random effects defined by (3.4). Thus, it is interesting to test whether the use of the PWIGLS algorithm for estimating the cross-sectional model parameters but without weighting the time series likelihood likewise controls the bias. Table 4 shows the results obtained for this case with the same samples as used for Tables 2 and 3. The estimators of the fixed vector coefficients $\beta_i = (\gamma_i', v_i')'$ are the same as in Table 3 and hence are not shown again. Notice that the estimators of Δ_{11}^* , Δ_{22}^* and σ_e^2 under Method 2 are also the same as the corresponding estimators in Table 3.

The interesting result revealed from Table 4 is that the estimators of A_{11} and A_{22} have now a non-negligible bias,

unlike the corresponding estimators in Table 3. This result can be explained as follows. Under the informative sampling scheme, the expectation of the random effects $u_{h,1,1}$ corresponding to households h in the sample is below zero, $E(u_{h,1,1} | h \in s) < 0$, and hence the initialization of the Kalman filter by the population expectation ($E u_{h,1,1} = 0$, Equation 4.4) yields biased estimators. On the other hand, by weighting the likelihood contributions L_h by the inverse of the sample selection probabilities, the proportions of likelihoods L_h corresponding to random effects that are below and above the model expectation is balanced to the population proportions and thus the use of the model expectation for the initialization process does not bias the estimation process. As noticed for the previous tables, the SD's of the unweighted estimators in Table 4 are much smaller than the SD's of the corresponding weighted estimators in Table 3.

Table 4

Means, Standard Deviations (SD) and *t*-Statistics of Estimators Under Two Estimation Methods. Informative Sampling, Weighted MLM, Unweighted Likelihood

Parameter	True Value	Method 1			Method 2		
		Mean	SD	<i>t</i> -statistic	Mean	SD	<i>t</i> -statistic
A_{11}	0.500	0.468	0.09	-3.477	0.453	0.10	-4.569
A_{22}	0.700	0.742	0.11	3.948	0.737	0.11	3.197
Δ_{11}^*	1.067	1.060	0.11	-0.598	1.040	0.17	-1.560
Δ_{22}^*	0.980	1.008	0.11	2.449	1.010	0.19	1.571
ρ	0.400	0.407	0.02	3.021	0.402	0.02	0.894
σ_e^2	0.298	0.298	0.01	1.013	0.294	0.01	-3.340

6. APPLICATION OF THE MODEL TO LFS DATA

We fitted the model defined by (3.1) and (3.4) to an empirical data set extracted from data collected by the Israel LFS for Jerusalem during the years 1990-1994. The data contain complete records for 567 individuals in 475 households, with each individual observed in four quarters according to the rotation pattern described before and used for the simulation study. Out of the 475 households, 385 have one individual record, 88 have 2 individual records and only 2 households have 3 individual records. The outcome variable is y = number of hours worked during the week preceding the interview, ($\bar{y} = 39.8$, $sd(y) = 14.8$; calculated over all individuals and all the quarters). The individual level auxiliary variables are x_1 = years of education, ($\bar{x}_1 = 13.4$, $sd(x_1) = 4.8$) and x_2 = gender, (41% females). The household level auxiliary variables are $z_1 = 1$ and z_2 = number of employed persons in the household ($\bar{z}_2 = 1.48$, $sd(z_2) = 0.56$).

We estimated the model parameters using the two methods described in section 4. The sampling weights attached to these data are very similar across households and individuals so that we only computed the unweighted

estimators. The LGLS algorithm produced negative variance estimates for Δ_{22}^* in some of the quarters and these estimates have been set to zero when averaging the variance estimates under Method 2. The quarterly estimates of the fixed model coefficients have not been averaged as they change significantly over the five years period.

The estimates computed by the two methods for the variances and autoregression coefficients are shown in Table 5 using the same notation as in the previous tables. The two sets of estimates are not very far except for the estimator of Δ_{22}^* which, has already mentioned was found to be negative in some of the separate IGLS runs. Note in this respect that for most of the households there is only a single individual record (see above), and that for almost all of these households $z_2 = 1$. This complicates the estimation process since for such households it is impossible to distinguish the first (individual) level effect from the two household effects, which are likewise confounded. (Note that the sum of the latter two variances is similar under the two methods.) As discussed below, the estimators in Table 5 are dominated by the observations obtained for households with two individual records.

Table 5
Estimates of Variances and Autoregression
Coefficients Under Two Estimation Methods.
LFS Data

Parameter	A_{11}	A_{22}	Δ_{11}^*	Δ_{22}^*	ρ	σ_e^2
Method 1	0.915	-0.606	73.88	2.541	0.242	102.306
Method 2	0.976	-0.548	56.88	14.753	0.448	101.001

Under the Israel LFS sampling design, each individual record consists of 4 observations taken in quarters 1, 2, 5 and 6, with quarter 1 defining the first calendar quarter t that the individual is in the sample. In order to assess the prediction power of the model, we computed for every individual record (h, j) the empirical innovations when predicting the adjusted values $r_{hjq} = (y_{hjq} - x'_{hjq}\hat{\gamma}_q - z'_{hjq}\hat{\gamma}_q)$ using the household data observed for the preceding quarters that the individual has been in the sample. Note that by subtracting the fixed effects from the original observations, the distribution of the adjusted values no longer depends on the calendar quarters. The innovation for quarter q is the corresponding prediction error which, by (3.1) is computed as $d_{hjq} = (r_{hjq} - z'_{hjq}\hat{\alpha}_{hjq|q-m} - \hat{e}_{hjq|q-m}) = r_{hjq} - (z'_{hq}, 1)' \hat{\alpha}_{q|q-m}$, $q = 2, 5, 6$ where $\hat{\alpha}_{q|q-m}$ is the predictor of the state vector $\alpha_q = (u'_{hq}, e'_{hjq})'$ using the data observed until quarter $q-m$, with $m = q-1$ for $q = 2, 6$ and $m = 3$ for $q = 5$. The predictor $\hat{\alpha}_{q|q-m}$ is obtained by application of the Kalman filter with the corresponding estimated parameters (see section 2.3 and Equations 3.5 and 4.3).

Table 6 shows the roots of the means of the square innovations (RMSI) by quarter and the number of household (HH) records, as obtained under the two estimation methods (using the parameter values displayed in Table 5).

For comparison, we also show the RMSI's of the innovations obtained by predicting the adjusted value for quarter q by the adjusted value in the preceding quarter. The "naive" predictor $\hat{r}_{hjq} = r_{hj, q-m}$ can be interpreted as being the optimal predictor under the simple random walk model $r_{hjq} = r_{hj, q-m} + \text{error}$. The means of the innovations $(\hat{r}_{hjq} - r_{hj, q-m})$ for $q = (2, 5, 6)$ are (0.68, 0.24, 0.301) for households with one record, (1.24, -1.20, 0.60) for households with two records and (4.02, -5.82, 7.68) for households with 3 records but recall that the latter means are based on only 2 households. The corresponding means of the empirical innovations computed under the model are smaller in absolute value in all the cases.

Table 6
Root Mean Square of Innovations by Number of Household
Records and Quarter Under Two Estimation Methods
and Naive Prediction. LFS Data

HH Records	1			2			3		
Quarter	2	5	6	2	5	6	2	5	6
Method 1	11.54	11.16	11.62	12.26	11.71	10.88	9.61	9.98	8.94
Method 2	11.71	11.16	11.49	12.10	11.48	10.91	9.30	9.78	7.90
Naive Pred.	14.00	11.92	13.60	14.71	15.12	13.47	7.50	13.32	11.29

The data analyzed in this section behave much more erratically than the data used for the simulation study generated under the model and we cannot claim that the model employed yields the best possible fit (see also below). Nonetheless, the values displayed in Table 6 illustrate some important features of the model. We mention first the generally much better performance of the model predictors compared to the naive predictor $\hat{r}_{hjq} = r_{hj, q-m}$ with the two estimation methods yielding similar RMSI's. The superiority of the model is explained by the fact that whereas the first order autocorrelations of the two random household effects used for the model predictions are high in absolute value (very high for the first component), the autocorrelations of the adjusted values (the "total" errors) are only of moderate size. The first order autocorrelations of the random components are the corresponding autoregression coefficients, see Table 5. The empirical autocorrelations of the adjusted values, $\text{Corr}(\hat{r}_{hjq}, \hat{r}_{hj, q-m})$, $q = 2, 5, 6$; $m = 1$ for $q = 2, 6$; $m = 3$ for $q = 5$ are correspondingly (0.46, 0.59, 0.51) for one record households, (0.48, 0.36, 0.45) for two record households and (0.92, 0.43, 0.63) for three record households (based on 6 individual records).

As already noted, the fact that most households have only one individual record introduces identifiability problems since for such households it is impossible to distinguish between the three random effects. Computation of the correlations $\text{Corr}(\hat{r}_{hjq}, \hat{r}_{hj, q-m})$, under the model using the parameter estimates in Table 5 shows a good fit to the correlations computed for two record households. This in turn illustrates that the estimators in Table 5 are

dominated by these observations and we conclude that the model fits best the observations obtained for the households with two records. Note, however, that the RMSI's obtained for the other household sizes are not higher than the RMSI's computed for the two record households (see also below). It is important to mention in this regard that if the data had been aggregated over all the individuals observed in a given calendar quarter, it would have been impossible to account for the random household effects, resulting in inferior predictions of the individual observations. See the discussion in the introduction. (Modelling the aggregate data is rather complicated in this case since the sample in each calendar quarter consists of 4 different panels as defined by the number of times that individuals are in the sample. This implies that the models holding for these panels are different, depending on the number of observations available for each panel.)

Other interesting results noted in Table 6 are that the RMSI's under the model are generally lower for $q = 6$ than for $q = 2$, as explained by the use of more observed data for the same individual in the prediction process (more observed data for estimating the random effects in the preceding quarter). Also, for $q = 6$ the RMSI's decrease as the number of household records increases, as explained by the use of data observed for other household members. Finally, the RMSI's for households with 3 records are much lower by use of the model than the RMSI's obtained for households with 1 and 2 records but we mention again that there are only 2 households with three records. The unexpected results in Table 6 are that for households with one record the RMSI's are somewhat larger for $q = 6$ than for $q = 5$ (note the relatively high and unexplained correlation of 0.59 between the adjusted values 3 quarters apart computed for these households), and that for $q = 2$ and $q = 5$ the RMSI's for households with 2 records are larger than the corresponding RMSI's for households with 1 record. With empirical data of relatively small size such anomalies are not unusual and they show up even more prominently with the naive predictor. (The fact that for a given number of household records the RMSI's by use of the model for $q = 5$ are of similar magnitude to the other RMSI's is reassuring given that the predictions in this case are 3 quarters ahead.)

7. CONCLUSIONS AND MODEL EXTENSIONS

The results of this paper illustrate that it is possible to fit time series models to longitudinal series of very short length and with missing observations. The model used in the present study is an extension of the standard two level linear model by which both the first and second level random effects evolve stochastically over time. This kind of model is suitable for modelling longitudinal measurements that are taken for hierarchical populations. Application of the PWIGLS algorithm combined with standard probability

weighting of the time series likelihood is shown to protect against the effects of informative sampling.

Multilevel models are often fitted to discrete data, in which case the models contain nonlinear components. In principle, the two-stage estimation method proposed in this paper can be applied in this case as well, although with very short longitudinal series the range of models that can be fitted is obviously limited. Moreover, a common procedure for estimating the unknown model parameters in the discrete case consists of linearizing the nonlinear components on each iteration of the IGLS around estimates obtained on the previous iteration, and then applying the standard IGLS for computing the revised estimates. See Goldstein (1995) for details. Thus, it seems feasible to extend the PWIGLS algorithm to the discrete case without major difficulties.

In this paper we have not considered variance estimation. This is no problem under the standard IGLS and Pfeffermann *et al.* (1998) propose simple variance estimators for the PWIGLS procedure. However, estimation of the variances of estimators obtained from maximization of the time series likelihood is more problematic because of two reasons. First, the possibly short length of the longitudinal series may no longer justify the use of the information matrix or permit stable estimation thereof, even with large number of second level units. Second, the MLM estimators are held fixed when maximizing the likelihood, implying that the MLE abstract from the sampling errors in the estimation of the MLM parameters. A possible solution to this problem is the use of re-sampling methods that allow to account for all sources of variation in the estimation process.

Finally, we mention an important application of the proposed model for the imputation of missing data. In a recent article, Pfeffermann and Nathan (forthcoming) illustrate the large reductions in the imputation variance that can be achieved under the model compared to the use of more standard imputation methods that ignore the common household effects.

REFERENCES

- BELCHER, J., HAMPTON, J.S., and TUNNICLIFFE WILSON, G. (1994). Parameterization of continuous time autoregressive models for irregularly sampled time series data. *Journal of the Royal Statistical Society, Series B*, 56, 141-155.
- BINDER, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.
- BINDER, D.A. (1998). Longitudinal surveys: why are these surveys different from all other surveys? *Survey Methodology*, 24, 101-108.
- BRYANT, J., and DAY, R. (1991). Empirical Bayes analysis for systems of mixed models with linked autocorrelated random effects. *Journal of the American Statistical Association*, 86, 1007-1012.

- CHI, E.M., and REINSEL, G.C. (1989). Models for longitudinal data with random effects and AR(1) errors. *Journal of the American Statistical Association*, 84, 452-459.
- DIGGLE, P.J., LIANG, K.Y., and ZEGER, S.L. (1994). *Analysis of Longitudinal Data*. Oxford: Clarendon Press.
- DUNCAN, G.J., and KALTON, G. (1987). Issues of design and analysis of surveys across time. *International Statistical Review*, 55, 97-117.
- GOLDSTEIN, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, 73, 43-56.
- GOLDSTEIN, H. (1995). *Multilevel Statistical Models* (2nd edition). New York: Halstead.
- GOLDSTEIN, H., HEALY, M.J.R., and RASBASH, J. (1994). Multilevel time series models with applications to repeated measures data. *Statistics in Medicine*, 13, 1643-1655.
- HARVEY, A.C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. New York: Cambridge University Press.
- HERRIOT, R.A., and KASPRZYK, D. (1984). The survey of income and program participation. *Proceedings of the Social Statistics Section, American Statistical Association*, 107-116.
- JONES, R.H., and ACKERSON, L.M. (1990). Serial correlation in unequally spaced longitudinal data. *Biometrika*, 77, 721-731.
- JONES, R.H., and BOADI-BOATING, F. (1991). Unequally spaced longitudinal data with AR(1) serial correlation. *Biometrics*, 47, 161-175.
- JONES, R.H., and VECCHIA, A.V. (1993). Fitting continuous ARMA models to unequally spaced spatial data. *Journal of the American Statistical Association*, 88, 947-954.
- JONES, R.H. (1993). *Longitudinal Data with Serial Correlation. A State-space Approach*. New York: Chapman and Hall.
- LAWLESS, J.F. (1999). Event History Analysis and Longitudinal Surveys. Paper presented at the Conference on Analysis of Survey Data, Southampton, United Kingdom.
- NATHAN, G. (1999). A Review of Sample Attrition and Representativeness in Three Longitudinal Surveys. GSS Methodology Series No. 13. London Office of National Statistics (ONS), United Kingdom.
- PFEFFERMANN, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61, 317-337.
- PFEFFERMANN, D. (1996). The use of sampling weights for survey data analysis. *Statistical Methods in Medical Research*, 5, 239-261.
- PFEFFERMANN, D., SKINNER, C.J., GOLDSTEIN, H., HOLMES, D.J., and RASBASH, J. (1998). Weighting for unequal selection probabilities in multilevel models (with discussion). *Journal of the Royal Statistical Society, Series B*, 60, 23-40.
- PFEFFERMANN, D., and NATHAN, G. (forthcoming). Imputation for wave nonresponse: existing methods and a times series approach. To appear in: *Survey Nonresponse*, (Eds. R.M. Groves, D. Dillman, J.L. Eltinge, and R.J.A. Little). New York: John Wiley and Sons.
- RAO, J.N.K. (1999). Some recent advances in model-based small area estimation. *Survey Methodology*, 25, 175-187.
- SALLAS, W.H., and HARVILLE, D. A. (1981). Best linear recursive estimation for mixed linear models. *Journal of the American Statistical Association*, 76, 860-869.
- SKINNER, C.J., HOLT, D., and SMITH, T.M.F. (Eds.) (1989). *Analysis of Complex Surveys*. Chichester: Wiley.
- SKINNER, C.J., and HOLMES, D. (1999). Random Effects Models for Longitudinal Survey Data. Paper presented at the Conference on Analysis of Survey Data, Southampton, United Kingdom.
- SURVEY RESEARCH CENTER (1984). User Guide to the Panel Study of Income Dynamics. Ann Arbor, Michigan: Inter-university Consortium for Political and Social Research.
- WEBBER, M. (1994). The survey of labor and income dynamics: lessons learned in testing. *Proceedings of the Annual Research Conference, US Bureau of the Census*, 85-99.
- ZIMMERMAN, D.L., and NUNEZ-ANTON, V. (1997). Structured antedependence models for longitudinal data. In: *Modelling Longitudinal and Spatially Correlated Data: Methods, Applications and Future Directions*, (Eds. T.G. Gregoire et al.). Lecture Notes in Statistics, 22. New York: Springer Verlag, 62-76.

A Conditional Mean Squared Error of Small Area Estimators

LOUIS-PAUL RIVEST and EVE BELMONTE¹

ABSTRACT

This paper suggests estimating the conditional mean squared error of small area estimators to evaluate their accuracy. This mean squared error is conditional in the sense that it measures the variability with respect to the sampling design for a particular realization of the smoothing model underlying the small area estimators. An unbiased estimator for the conditional mean squared error is easily constructed using Stein's Lemma for the expectation of normal random variables. This estimator can be calculated for any shrinking strategy; composite and empirical Bayes estimators are considered in this work. It can be calculated when the small area estimators are benchmarked to coincide with direct estimators at high level of aggregation. It can accommodate skewness in the data and estimated variances. The conditional mean squared error estimator does not rely on any smoothing model. The price to pay for this property is a high variance; the new estimator is unstable under heavy shrinking. In these situations, it still provides useful diagnostic information about the shrinking model. It can also be seen as a building block for estimators of unconditional mean squared errors such as Prasad and Rao's (1990). Examples dealing with the estimation of the under-coverage in the Canadian Census illustrate the application of this new estimator.

KEY WORDS: Census under-coverage; Diagnostics; Empirical Bayes estimation; Estimated variances; Skewness; Stein's lemma; Survey sampling.

1. INTRODUCTION

In survey sampling, the need to develop accurate methods of estimation for small areas poses challenging statistical problems. For small areas, direct survey estimates have too large a variance to be reliable. Small area techniques "improve" direct estimates by shrinking them towards model based smoothed values. Simple shrinking estimators are proposed by Purcell and Kish (1979). In a pioneering paper, Fay and Herriot (1979) demonstrate that this can lead to interesting gains in precision. The review papers of Ghosh and Rao (1994) and of Singh, Gambino and Mantel (1994) provide convincing evidence of the vitality of this area.

The estimation of the errors in small area estimation is receiving an increasing attention, see Singh, Stukel and Pfeiffermann (1998) and Booth and Hobert (1998). This paper suggests estimating the conditional mean squared errors of small area estimators. The conditional mean squared error can be estimated for all shrinking strategies, either empirical Bayes or decision theoretic (Purcell and Kish 1979). Other mean squared errors, such as Prasad and Rao's (1990), and Singh, Stukel and Pfeiffermann (1998) frequentist proposals measure the variability with respect to both, the sampling design and the smoothing model. The mean squared error of this paper is conditional in the sense that it measures variability with respect to the sampling design for a particular realization of the smoothing model. This feature is attractive since the conditional estimator reflects the conditions under which the survey has been carried out (see Särndal, Swensson, and Wretman 1992, ch.

7). The drawback of this property is a high variability. In some instances, the proposed estimator is too variable for practical use.

When shrinking is important, the conditional mean squared error estimators are highly unstable. An unconditional assessment of the precision of small area estimators must be used. In this situation, the conditional estimator proposed in this paper still provides some useful information. It can be looked at as a diagnostic for comparing smoothing models. It can also be a building block for constructing Monte Carlo estimates of unconditional mean squared errors in situations where closed form formulas, such as Prasad and Rao's (1990), are not available.

The assessment of the accuracy of estimators for the under-coverage, at the provincial and sub-provincial levels, of the Canadian Census motivated this work. Alternatives to the direct estimates for provincial under-coverage are discussed by Royce (1992) and Rivest (1995). Dick (1995) applies empirical Bayes methods to sub-provincial under-coverage estimates. These two examples are treated in section 5.

An estimator of the conditional mean squared error is presented in section 2. Its construction relies on the multivariate version of Stein's Lemma for the expectation of normal deviates. Section 3 suggests changes to the conditional estimator to accommodate skewness in the distribution of the direct estimators and estimated variances. Section 4 discusses the application of the new estimator to empirical Bayes estimators. Its relationship with Prasad and Rao (1990) prediction variance is highlighted. Examples are treated in section 5.

¹ Louis-Paul Rivest and Eve Belmonte, Département de mathématiques et de statistique, Université de Laval, Ste-Foy, Québec, Canada, G1K 7P4.

2. A CONDITIONAL MEAN SQUARED ERROR ESTIMATOR

Suppose that there are n small areas and let $\mu = (\mu_1, \dots, \mu_n)'$ denote the unknown population characteristics for these small areas. The direct survey estimates for the n small areas are $y = (y_1, \dots, y_n)'$ where the distribution of y is $N_n(\mu, \Sigma)$, a n -variate normal distribution with mean vector μ and known variance-covariance matrix Σ . As pointed out by Ghosh and Rao (1994), the normality assumption is likely to hold for many surveys since direct survey estimates are usually functions of sums of variables. The $n \times n$ matrix Σ is a design based measure of precision for y . For the time being, this matrix is assumed to be known. This assumption is relaxed in section 3.2. The uncertainty in y comes from the random selection of the sampling units. Subscript S , for sampling design, denotes expectations taken with respect to the distribution of y .

In a typical application of small area techniques, one has,

$$y_i = \frac{\sum_j w_{ij} y_{ij}}{\sum_j w_{ij}}$$

where y_{ij} is the y -value for the j -th sample unit in small area i , w_{ij} is its sampling weight and the sum is over all the sample units in small area i . In many instances, the variance covariance matrix Σ is diagonal; its (i, i) term, is $\sigma_{ii} = \text{Var}_S(y_i)$; when they are non null, the off diagonal elements of Σ are denoted by σ_{ij} , $i, j = 1, \dots, n$.

Several methods have been proposed to improve the accuracy of direct survey estimators. They involve shrinking y_i towards some indirect estimator of μ_i . The resulting estimators can be written as

$$\hat{\mu}_i = y_i + g_i(y_1, \dots, y_n), \quad i = 1, \dots, n \quad (1)$$

where the g_i 's are functions depending on the shrinking strategy.

In vector form, one can write (1) as $\hat{\mu} = y + g(y)$ where g , whose i -th component is equal to g_i , is a function defined from R^n to R^n . We assume that for each i , the right partial derivative and the left partial derivative of g_i with respect to y_j exists for any y in R^n . When they are equal, $\partial g_i(y)/\partial y_j$ denotes the common value; if they differ $\partial g_i(y)/\partial y_j$ is the average between the two values. The component of $g(y)$ and their partial derivatives are assumed to have finite variances. A conditional assessment of the precision of $\hat{\mu}$ as an estimator for μ is given by the matrix of the mean product errors which is given by

$$E_S\{(\hat{\mu} - \mu)(\hat{\mu} - \mu)'\} = \Sigma + E_S\{(y - \mu)g(y)'\} + E_S\{g(y)(y - \mu)'\} + E_S\{g(y)g(y)'\}$$

On the right hand side of this equality, the only quantities for which there are no obvious estimators are $E_S\{(y - \mu)g(y)'\}$ and $E_S\{g(y)(y - \mu)'\}$. Their evaluations

are eased by the following result which is a multivariate extension of Stein's lemma (Stein 1981). Its proof is given in the appendix together with the proofs for Propositions 2, 3, and 4.

PROPOSITION 1: Let y be a $N_n(\mu, \Sigma)$ random vector then,

$$E_S\{(y - \mu)g(y)'\} = \Sigma E_S\{\nabla g(y)\},$$

where $\nabla g(y)$ is an $n \times n$ matrix whose (i, j) -th element is given by $g_j^i(y) = \partial g_i(y)/\partial y_j$.

Now according to Proposition 1, $\Sigma \nabla g(y)$ is an unbiased estimator for $E_S\{(y - \mu)g(y)'\}$. Thus the conditional estimator (index "c" stands for conditional) for the matrix of the mean product errors is given by

$$\text{mpe}_c(\hat{\mu}) = \Sigma + \Sigma \nabla g(y) + \nabla g(y)' \Sigma + g(y)g(y)'. \quad (2)$$

The diagonal terms of (2) can be negative. Since they estimate mean squared errors, a better estimator for the mean squared error of $\hat{\mu}_i$ is

$$\text{mse}_c^*(\hat{\mu}_i) = \max\left\{0, \sigma_{ii} + \sum_j \sigma_{ij} \{g_j^i(y) + g_j^i(y)\} + g_i(y)^2\right\}.$$

It generalizes an estimator proposed by Bilodeau and Srivastava (1988) for James-Stein estimator, and by Robert (1992 p. 292) for empirical Bayes estimators. When the y_i 's are independent, with $\sigma_{ij} = 0$ when $i \neq j$, then

$$\text{mse}_c(\hat{\mu}_i) = \sigma_{ii} + 2\sigma_{ii} \frac{\partial g_i(y)}{\partial y_i} + g_i(y)^2, \quad (3)$$

and $\text{mse}_c^*(\hat{\mu}_i) = \max\{\text{mse}_c(\hat{\mu}_i), 0\}$.

Kott's (1989) small area estimator has $g_i(y) = \hat{\alpha}_i(\hat{\gamma}_i - y_i)$, where $\hat{\gamma}_i$ is a measure of location for the y 's and $\hat{\alpha}_i$ is a smoothing parameter. These two statistics involve variance estimates calculated at the "unit" level, that is using the y_{ij} 's. Kott's (1989) conditional mean squared error is

$$v(\hat{\mu}_i) = \sigma_{ii}(1 - 2\hat{\alpha}) + (\hat{\alpha}_i(y_i - \hat{\gamma}_i))^2.$$

This is equal to (3) when both $(d/dy_i)\hat{\alpha}_i$ and $(d/dy_i)\hat{\gamma}_i$ are null. Thus Kott's (1989) estimator for the conditional mean squared error does not account for the estimation for the variance components. This may account for the biases that it exhibited in the simulations reported by Prasad and Rao (1999).

The estimates mse_c and mpe_c can be evaluated numerically by taking

$$\frac{\partial g_i(y)}{\partial y_j} = \frac{g_i(y_1, \dots, y_{j-1}, y_j + \epsilon, y_{j+1}, \dots, y_n) - g_i(y_1, \dots, y_{j-1}, y_j - \epsilon, y_{j+1}, \dots, y_n)}{2\epsilon}$$

where ϵ is a small positive number. Thus mse_c and mpe_c can be calculated in all circumstances, even when g has no explicit form.

To illustrate the flexibility of the conditional estimator, consider $\hat{\mu}^* = \hat{\mu}(\Sigma y_i) / (\Sigma \hat{\mu}_i)$, an estimator bench-marked to agree with the direct estimator for the y -total. One has $\hat{\mu}^* = y + g^*(y)$ where

$$g^*(y) = \frac{\sum y_i}{\sum \hat{\mu}_i} g(y) + \left(\frac{\sum y_i}{\sum \hat{\mu}_i} - 1 \right) y.$$

It might be difficult to derive an analytical formula for $\text{mpe}_c(\hat{\mu}^*)$, however this expression is easily evaluated using numerical derivatives. Modifications of the conditional estimator to account for non-normality in the y_i 's and for estimated variances σ_{ii} are given next.

3. SENSITIVITY ANALYSIS

In many surveys, especially those in the business sector, the study variables are skewed. Some of this skewness might still be left in the direct estimators y_i . This section suggests a correction to the conditional mean squared error to account for skewness in the distribution of y . It also proposes ways to account for the estimation of the variances σ_{ii} in the mean squared error calculations.

In practice the variances σ_{ii} are estimated. Several authors (Dick 1995; Hogan 1992) smooth the variances before calculating the small area estimates. They then consider the smoothed variances as the true variances in the small area calculations. Section 3.2 gives a condition under which replacing the estimated variances by their smoothed values yields unbiased mean squared error estimators. It also considers situations where the sampling variances are estimated with random groups (Wolter 1985 ch.2). This method consists in carrying a certain number, say k , of replications of the survey design. This yields, for each i , k estimates of μ_i ; $\hat{\sigma}_{ii}$ is then equal to the sampling variance of these k estimates divided by k . Assuming that these k estimates are normally distributed, one can consider that, suitably normalized, the distribution of $\hat{\sigma}_{ii}$ is chi-squared with $k - 1$ degrees of freedom. A conditional mean squared error, adjusted for variances estimated with random groups, is proposed in this section. To keep the discussion simple, we assume in this section that Σ is a diagonal matrix; in other words the y_i 's are assumed to be independent random variables.

3.1 Non-Normality in the Distribution of y_i

In many applications of small area estimation, the distributions of the y_i 's are not exactly normal. A simple adjustment to (3) is proposed to deal with asymmetry in the distribution of the y_i 's.

Suppose that the skewness of y_i , $\rho_i = E_S\{(y_i - \mu_i)^3\} / \sigma_{ii}^{3/2}$ is small and non-zero. A first order Edgeworth series for the distribution of y_i is given by (see for instance Reid 1991):

$$f(t) = \frac{\exp\{- (t - \mu_i)^2 / (2\sigma_{ii})\}}{\sqrt{2\sigma_{ii}\pi}} \times \left[1 + \frac{\rho_i}{6} \left\{ \left(\frac{t - \mu_i}{\sqrt{\sigma_{ii}}} \right)^3 - 3 \left(\frac{t - \mu_i}{\sqrt{\sigma_{ii}}} \right) \right\} \right].$$

Such an expansion is used to correct for skewness in the direct estimators (Barndorff-Nielsen and Cox 1989, remark 2 p. 92). Expansions involving additional terms are used for correcting for both skewness and kurtosis; they will not be considered in this section. The evaluation of $E\{(y_i - \mu_i)g_i(y)\}$ under f , needed for the construction of the conditional mean squared error estimator, is given in Proposition 2.

PROPOSITION 2: When y_i distributed according to $f(t)$, as ρ_i tends to 0.

$$E_S\{(y_i - \mu_i)g_i(y)\} = \sigma_{ii} E_S \left\{ \frac{\partial g_i(y)}{\partial y_i} \right\} + \frac{\sigma_{ii}^{3/2} \rho_i}{2} E_S \left\{ \frac{\partial^2 g_i(y)}{\partial y_i^2} \right\} + O(\rho_i).$$

A mean squared error estimator corrected for asymmetry is therefore given by $\text{mse}_c^*(\hat{\mu}_i) = \max\{0, \text{mse}_c(\hat{\mu}_i)\}$ where

$$\text{mse}_c(\hat{\mu}_i) = \sigma_{ii} + 2\sigma_{ii} \frac{\partial g_i(y)}{\partial y_i} + \sigma_{ii}^{3/2} \rho_i \frac{\partial^2 g_i(y)}{\partial y_i^2} + g_i(y)^2.$$

In practice, it might be difficult to find individual skewness coefficients ρ_i for each i . A better strategy might be to combine all the data points to come up with a common ρ -value.

3.2 Estimated Variances

Consider first a survey where the $\hat{\sigma}_{ii}$'s are estimated using k random groups. Assuming normality, one can consider that $\{(k - 1)\hat{\sigma}_{ii} / \sigma_{ii}; i = 1, \dots, n\}$ is a sequence of independent χ_{k-1}^2 random variables which is independent of y . Evaluating the conditional mean squared error (3) with variance estimates $\hat{\sigma}_{ii}$ yields potentially biased estimators, since $g_i(y)$ and its derivatives depend on $\hat{\sigma}_{ii}$. The potential bias can be expressed as

$$2E\left\{\hat{\sigma}_{ii} \frac{\partial g_i(y)}{\partial y_i}\right\} - 2\sigma_{ii} E\left\{\frac{\partial g_i(y)}{\partial y_i}\right\} \quad (4)$$

As shown in the Appendix, this bias is $O(1/k)$. The next proposition suggests a small change to (3) that reduces its bias (4).

PROPOSITION 3: Replacing $\hat{\sigma}_{ii}$ by $(k - 1)\hat{\sigma}_{ii} / (k + 1)$ in the evaluation of $\partial g_i(y) / \partial y_i$ for calculating the mean squared error estimator (3) yields an estimator with an $O(1/k^2)$ bias.

The correction factor $(k-1)/(k+1)$ has been proposed in a different context by Scott and Smith (1971). Other methods are available for correcting the bias for estimating variances, depending on the way in which σ_{ii} is estimated. For instance if the $\hat{\sigma}_{ii}$ are independent $N\{\sigma_{ii}, \text{var}(\hat{\sigma}_{ii})\}$ random variables distributed independently of y_i , then by Stein's lemma, (4) is equal to $2\text{var}(\hat{\sigma}_{ii})E\{\partial^2 g_i(y)/\partial y_i \partial \hat{\sigma}_{ii}\}$.

Suppose now that the variances are estimated, not necessarily with random groups. In surveys, such as those considered in Dick (1995) and Hogan (1992), explanatory variables are available to model estimated variances. Small area estimators are then calculated with the predicted variances $\tilde{\sigma}_{ii}$ under the smoothing model; this means that $\tilde{\sigma}_{ii}$ enters in the calculation of g_i in (1). Considering (4), the mean squared error estimated with the smoothed variance,

$$\tilde{\sigma}_{ii} + 2\tilde{\sigma}_{ii} \frac{\partial g_i(y)}{\partial y_i} + g_i(y)^2$$

is unbiased provided that

$$2E\left\{\left(\tilde{\sigma}_{ii} - \sigma_{ii}\right) \frac{\partial g_i(y)}{\partial y_i}\right\} = 0.$$

When $g_i(y)$ is calculated with smoothed variances, (4) should be small; the above condition holds provided that

$$E_V\left\{\left(\hat{\sigma}_{ii} - \tilde{\sigma}_{ii}\right) \frac{\partial g_i(y)}{\partial y_i}\right\} = 0, \quad (5)$$

where index V refers to the model for smoothing the variances. One can easily test whether this condition holds by calculating the correlation between the variance residuals and the partial derivatives of the functions g_i . Since, as shown in Proposition 5 of the next section, unconditional mean squared errors can be derived as expectations of $\text{mse}_c(\hat{\mu}_i)$ testing whether (5) is true is relevant even when unconditional measures of accuracy, such as Prasad and Rao's, are calculated. Indeed, replacing variances by their predicted values biases the mean squared error estimators, conditional or unconditional, when (5) is violated.

4. MEAN SQUARED ERROR ESTIMATION FOR EMPIRICAL BAYES ESTIMATORS

4.1 Model Construction

This section assumes that the y_i 's are independent, i.e., that Σ is diagonal. In an empirical Bayes setting, the model (M) for smoothing direct estimators expresses the parameters μ_i 's as random variables whose distributions depend on a p -variate auxiliary variable x_i (Maritz and Lwin 1989, chapter 3),

$$\mu_i = x_i' \beta + v_i, \quad (6)$$

where β is a $p \times 1$ vector of unknown regression parameters and the v_i 's are independent random variables with mean 0 and variance σ_v^2 . Often the v_i 's are assumed to be normally distributed; the marginal distribution of y_i , with respect to both the sampling design S and the smoothing model M , is then $N(x_i' \beta, \sigma_{ii} + \sigma_v^2)$. The empirical Bayes estimators are obtained by shrinking the direct estimators y_i towards their predicted values under (6).

The extent of the shrinking depends on estimators of the parameters of (6) calculated from the marginal distribution of y_i . Several methods are available for parameter estimation (Cressie 1992). A popular estimator for σ_v^2 (see Lahiri and Rao (1995)) is

$$\hat{\sigma}_v^2 = \max \left[0, (n-p)^{-1} \left\{ \sum_{i=1}^n (y_i - x_i' \hat{\beta})^2 - \sum_{i=1}^n \sigma_{ii} (1 - h_{ii}) \right\} \right]$$

where $\hat{\beta} = (X'X)^{-1}X'y$, $h_{ii} = x_i'(X'X)^{-1}x_i$, and $X = (x_1, \dots, x_n)'$. The weighted least squares estimator of β is

$$\hat{\beta}_w = \hat{A}^{-1} \sum_{i=1}^n \frac{x_i y_i}{(\hat{\sigma}_v^2 + \sigma_{ii})},$$

where

$$\hat{A} = \sum_{i=1}^n \frac{x_i x_i'}{(\hat{\sigma}_v^2 + \sigma_{ii})}.$$

The empirical Bayes estimator for μ_i is then

$$\hat{\mu}_i = x_i' \hat{\beta}_w + \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \sigma_{ii}} (y_i - x_i' \hat{\beta}_w) = y_i - \frac{\sigma_{ii}}{\hat{\sigma}_v^2 + \sigma_{ii}} (y_i - x_i' \hat{\beta}_w). \quad (7)$$

Thus for empirical Bayes estimators, one has

$$g_i(y) = - \frac{\sigma_{ii}}{\hat{\sigma}_v^2 + \sigma_{ii}} (y_i - x_i' \hat{\beta}_w).$$

4.2 The Conditional Mean Squared Error Estimator

An explicit form for (3) can be obtained from the following formula for the derivative of the functions g_i for empirical Bayes estimators,

$$\frac{\partial g_i(y)}{\partial y_i} = \frac{\partial \hat{\sigma}_v^2}{\partial y_i} \frac{\partial g_i(y)}{\partial \hat{\sigma}_v^2} - \frac{\sigma_{ii}}{\hat{\sigma}_v^2 + \sigma_{ii}} \left\{ 1 - \frac{x_i' \hat{A}^{-1} x_i}{(\hat{\sigma}_v^2 + \sigma_{ii})} \right\}, \quad (8)$$

The partial derivatives appearing in (8) can be evaluated using standard methods. They are given by

$$\frac{\partial \hat{\sigma}_v^2}{\partial y_i} = \frac{2}{(n-p)} (y_i - x_i' \hat{\beta}),$$

and

$$\frac{\partial g_i(y)}{\partial \hat{\sigma}_v^2} = \frac{\sigma_{ii}}{(\hat{\sigma}_v^2 + \sigma_{ii})^2} (y_i - x_i' \hat{\beta}_w) + \frac{\sigma_{ii}}{(\hat{\sigma}_v^2 + \sigma_{ii})} x_i' \frac{\partial \hat{\beta}_w}{\partial \hat{\sigma}_v^2},$$

where

$$\frac{\partial \hat{\beta}_w}{\partial \hat{\sigma}_v^2} = -\hat{A}^{-1} \sum_1^n \frac{x_k (y_k - x_k' \hat{\beta}_w)}{(\hat{\sigma}_v^2 + \sigma_{kk})^2}.$$

From (8), one has a closed form expression for $\text{mse}_c(\hat{\mu}_i)$. This statistic is an estimator of mean squared error for the empirical Bayes estimator for the i -th small area with respect to the sampling design only. It is valid for any sample size n ; it relies on the sole assumption that the direct estimators y_i are normally distributed. When $\hat{\sigma}_v^2 = 0$, $\hat{\mu}_i = x_i' \hat{\beta}_w$ and the derivatives in (8) simplify substantially. Since $\partial \hat{\sigma}_v^2 / \partial y_i = 0$, one has

$$\text{mse}_c(\hat{\mu}_i) = (y_i - x_i' \hat{\beta}_w)^2 - \sigma_{ii} + 2x_i' \hat{A}^{-1} x_i.$$

The properties of the conditional mean squared error estimator are best investigated in the simple situation where all the parameters of the smoothing model are assumed to be known. In this situation, $\partial g_i(y) / \partial y_i = -\sigma_{ii} / (\sigma_{ii} + \sigma_v^2)$ and the conditional mean squared error estimator is equal to $\text{mse}_c^*(\hat{\mu}_i) = \max \{(\text{mse}_c(\hat{\mu}_i), 0)\}$ where

$$\text{mse}_c(\hat{\mu}_i) = \frac{\sigma_{ii} \sigma_v^2}{\sigma_{ii} + \sigma_v^2} + \left(\frac{\sigma_{ii}}{\sigma_{ii} + \sigma_v^2} \right)^2 \{ (y_i - x_i' \beta)^2 - \sigma_{ii} - \sigma_v^2 \}.$$

The model based alternative to this estimator is the posterior variance, $\sigma_{ii} \sigma_v^2 / (\sigma_{ii} + \sigma_v^2)$, which coincides with $E_M[E_S\{\text{mse}_c(\hat{\mu}_i)\}]$. This estimator is a special case of Prasad and Rao (1990) estimator and is denoted $\text{mse}_{\text{PR}}(\hat{\mu}_i)$. Estimator $\text{mse}_c^*(\hat{\mu}_i)$ is highly variable when σ_v^2 is small. Indeed, when σ_v^2 is close to 0, about 50% of the conditional mean squared error estimates are null. To further compare the 2 mean squared error estimators, conditional and unconditional, observe that when all the parameters of the smoothing model are known, the conditional mean squared error of $\hat{\mu}_i$ is

$$E_S\{\text{mse}_c(\hat{\mu}_i)\} = \frac{\sigma_{ii} \sigma_v^2}{\sigma_{ii} + \sigma_v^2} + \left(\frac{\sigma_{ii}}{\sigma_{ii} + \sigma_v^2} \right)^2 \{ (\mu_i - x_i' \beta)^2 - \sigma_v^2 \}.$$

The next proposition compares the average mean squared errors of the estimators, conditional or unconditional, of $E_S\{\text{mse}_c(\hat{\mu}_i)\}$.

PROPOSITION 4: When $\sigma_{ii} = \sigma^2$, for $i = 1, \dots, n$ and when the small area means are μ_i 's are drawn using (6), the efficiency of the posterior variance with respect to the conditional mean squared error estimator for estimating the conditional mean squared error is

$$\frac{E_M\left[\sum \text{MSE}_S\{\text{mse}_c(\hat{\mu}_i)\} / n\right]}{E_M\left[\sum \text{MSE}_S\{\text{mse}_{\text{PR}}(\hat{\mu}_i)\} / n\right]} = \frac{\sigma^4 + 2\sigma^2 \sigma_v^2}{\sigma_v^4}$$

where $\text{MSE}_S(\cdot)$ denote a mean squared error taken with respect to the distribution of the y_i 's which are independent $N(\mu_i, \sigma^2)$ random variables.

The above efficiency is larger than 1 provided that $\sigma_v^2 / \sigma^2 < 2.41$. Proposition 4 shows under heavy shrinking, the unconditional mean squared error estimator is a better estimator of the conditional mean squared error than the conditional estimator. This surprising result is caused by the large variance of the conditional estimator; when shrinking is extensive, it is a poor estimator.

In some situations, such as that consider in section 5.1, shrinking is light and the use of the conditional mean squared error estimator is appropriate. The conditional efficiency of $\hat{\mu}_i$ with respect to the direct estimator y_i is given by $\sigma_{ii} / \text{mse}_c^*(\hat{\mu}_i)$. This is larger than one provided that $(y_i - x_i' \beta)^2 / (\sigma_{ii} + \sigma_v^2) < 2$. Assuming that the smoothing model holds true, conditional efficiencies less than 1 can be expected for approximately 16% ($= P[N(0,1)^2 < 2]$) of the small area estimators. This percentage should be higher if the smoothing model is deficient. Conditional efficiencies less than 1 occur in small areas having large residuals. On the other hand, the unconditional efficiencies, calculated with the posterior variance are, in this situation, less than 1 for all small areas. This shows that it is practically impossible for all the conditional efficiencies to be less 1; this had already been noted by Rao and Shinozaki (1978) for James-Stein estimators.

Many of the observations made in the unrealistic situation where all the parameters are known also apply when parameters are estimated. The unconditional alternative to the conditional mean squared error estimator is Prasad and Rao's (1990) estimator,

$$\text{mse}_{\text{PR}}(\hat{\mu}_i) = \frac{\sigma_{ii} \hat{\sigma}_v^2}{\sigma_{ii} + \hat{\sigma}_v^2} + \frac{\sigma_{ii}^2 x_i' \hat{A}^{-1} x_i}{(\sigma_{ii} + \hat{\sigma}_v^2)^2} + 2 \frac{\sigma_{ii}^2 \widehat{\text{Var}}(\hat{\sigma}_v^2)}{(\sigma_{ii} + \hat{\sigma}_v^2)^3}, \quad (9)$$

where $\widehat{\text{Var}}(\hat{\sigma}_v^2) = 2 \sum (\hat{\sigma}_{ii} + \hat{\sigma}_v^2)^2 / n^2$. To investigate the extent to which Proposition 4 holds when parameters are estimated, a small Monte Carlo study was carried out along the lines of the approach ii) simulation study of Prasad and Rao (1999). In all the simulations, $n = 30$ and $\sigma_{ii} = 1$, for $i = 1, \dots, n$. The smoothing model (6) was $\mu_i = \mu + v_i$ and various values of σ_v^2 were used. The results reported in Table 1 are based on $m = 5000$ Monte Carlo replications.

The simulations used 5 sets of μ_i -values whose variances are reported in Table 1. For each set, y_i was simulated repeatedly as a $N(\mu_i, 1)$ random variable, $i = 1, \dots, n$. The empirical Bayes estimate $\hat{\mu}_i$ was calculated for each small area and the mean squared error for small area i was calculated as $\text{MSE}_i = \sum^* (\hat{\mu}_i - \mu_i)^2 / m$ where \sum^* denotes the sum on the m Monte Carlo replications. The efficiency of the empirical Bayes estimator for small area i is $1 / \text{MSE}_i$. The mean and the median of the $n = 30$ small area efficiencies are given in Table 1. The 2 mean squared errors, conditional and unconditional, were calculated for

each small area in the m Monte Carlo replications; from (9), $\text{mse}_{\text{PR}}(\hat{\mu}_i) = (\hat{\sigma}_v^2 + 5/n) / (1 + \hat{\sigma}_v^2)$ for each small area. Table 1 presents the mean and the median of their absolute relative biases, defined as $|\sum^* (\text{mse}_i(\hat{\mu}_i) - \text{MSE}_i)| / (m\text{MSE}_i)$, and of their coefficients of variation which are equal to $(\sum^* (\text{mse}_i(\hat{\mu}_i) - \text{MSE}_i)^2 / m)^{1/2} / \text{MSE}_i$.

Table 1

Relative Efficiency of the Empirical Bayes Estimators (RE), Absolute Relative Bias (RB) and Coefficient of Variation (CV) of two MSE Estimators ($n = 30$). All Results are Expressed in Percentage

$\sum (\mu_i - \bar{\mu})^2 / 29$		RE%	RB _c %	RB _{PR} %	CV _c %	CV _{PR} %
1.3	mean	212	1	47	97	51
	median	214	1	40	100	43
2.53	mean	149	2	30	37	31
	median	163	2	31	37	32
3.7	mean	129	2	20	23	20
	median	133	1	21	24	21
4.24	mean	125	2	19	19	20
	median	131	1	22	20	22
4.93	mean	122	1	17	15	18
	median	133	1	17	13	17

As shown in section 2, $\text{mse}_c(\hat{\mu}_i)$ is unbiased; the biases reported in Table 1 are caused by Monte Carlo errors. When $n = 30$, the condition $\hat{\sigma}_v^2 / \sigma^2 > 2.4$ derived in Proposition 4 for the conditional estimator to improve on the unconditional estimator is not sufficient; the stronger condition $\hat{\sigma}_v^2 / \sigma^2 > 4$ is needed. Noteworthy is the fact that in Table 1, for $\sum (\mu_i - \bar{\mu})^2 / 29 > 2.5$, the CV of $\text{mse}_{\text{PR}}(\hat{\mu}_i)$ is only bias. Table 1 confirms that, when $\hat{\sigma}_v^2$ is of the same order of magnitude as σ_{ii} or smaller, the squared residual dominates the distribution of the conditional mean squared error estimator; in such cases Prasad and Rao (1990) unconditional estimator is a better estimator of conditional mean squared error. Even in situations when $\text{mse}_c(\hat{\mu}_i)$ cannot be recommended as an estimator for the conditional mean squared error, it still provides interesting diagnostic information: changes in the conditional estimators give a basis for comparing two smoothing models. This is illustrated in section 5.2.

4.3 Conditional Mean Squared Error and Prediction Variance

This section explores the relationship between the conditional mean squared error proposed in this paper and the prediction variance which is an unconditional measure of accuracy. Using the rotation of (6), the prediction variance is $\text{MSE}(\hat{\mu}_i) = E_M[E_S\{(\hat{\mu}_i - x_i'\beta - v_i)^2\}]$. From the construction of presented in section 2, one has

$$E_S\{\text{mse}_c(\hat{\mu}_i)\} = E_S\{(\hat{\mu}_i - x_i'\beta - v_i)^2\}.$$

Thus we have the following result,

PROPOSITION 5: The conditional mean squared error of empirical Bayes small area estimators satisfies,

$$E_M[E_S\{\text{mse}_c(\hat{\mu}_i)\}] = \text{MSE}(\hat{\mu}_i),$$

where $\text{MSE}(\hat{\mu}_i)$ is the unconditional prediction variance.

Proposition 5 shows that $\text{mse}_c(\hat{\mu}_i)$ can be looked at as an intermediate step in the evaluation of the unconditional mean squared error of $\hat{\mu}_i$. Consider for instance the calculation of Prasad and Rao (1990) $o(1/n)$ approximation to $\text{MSE}(\hat{\mu}_i)$,

$$\text{MSE}_{\text{PR}}(\hat{\mu}_i) = \frac{\sigma_{ii}\sigma_v^2}{\sigma_{ii} + \sigma_v^2} + \frac{\sigma_{ii}^2 x_i' A^{-1} x_i}{(\sigma_{ii} + \sigma_v^2)^2} + \frac{\sigma_{ii}^2 \text{Var}(\hat{\sigma}_v^2)}{(\sigma_{ii} + \sigma_v^2)^3},$$

where $\text{Var}(\hat{\sigma}_v^2) = 2 \sum (\sigma_{ii} + \sigma_v^2)^2 / n^2$. The standard derivation, as reviewed in section 3.2 of Singh, Stukel, and Pfeffermann (1998), is based on Kackar and Harville (1984). An alternative derivation, presented in Belmonte (1998, 1999), is to take the expectation of $\text{mse}_c(\hat{\mu}_i)$, obtained using (8), with respect to the marginal distribution of the y_i 's, which are independent $N(x_i'\beta, \sigma_{ii} + \sigma_v^2)$ deviates and to retain only the higher order terms.

Proposition 5 holds in situations where the small area estimators are bench-marked, or where corrections suggested in section 3 are implemented. These are cases for which there are no closed form formulas for the prediction variances. Proposition 4 suggests a simple method for constructing unconditional Monte Carlo estimates. It suffices to generate a large number of replicates of $\{y_i, i = 1, \dots, n\}$ where y_i follows a $N(x_i'\hat{\beta}_w, \hat{\sigma}_v^2 + \sigma_{ii})$ and to calculate $\text{mse}_c(\hat{\mu}_i)$ for each one. Averaging the $\text{mse}_c(\hat{\mu}_i)$'s gives a plug-in unconditional prediction variance, equal to the MSE of Proposition 4 evaluated at estimates $\hat{\beta}_w$, $\hat{\sigma}_v^2$ of the unknown parameters. Unfortunately, this estimate is biased (this is a first order estimate in the terminology of Singh, Stukel and Pfeffermann (1998)). For the empirical Bayes estimator given by (7), according to (9) the bias of the Monte Carlo estimate derived from Proposition 4 is $-\sigma_{ii}^2 \text{Var}(\hat{\sigma}_v^2) / (\sigma_{ii} + \hat{\sigma}_v^2)^3$. Further work is needed for constructing, using Proposition 4, unbiased unconditional prediction variance estimators.

5. ESTIMATING THE UNDER-COVERAGE IN THE 1991 CANADIAN CENSUS

In 1991, the under-coverage of the Canadian Census was estimated using two surveys, the Over-coverage Study, which estimates the number of persons double counted or erroneously counted in the Census and the Reverse Record Check (Burgess 1988) for the persons missed in the Census. Combining these figures gives estimates of the under-coverage of the Census. This section investigates several estimators of census under-coverage.

5.1 Provincial Estimations

The 1991 under-coverage rates for the ten Canadian provinces and the two territories with their coefficients of variation, expressed in percentage, are given in Table 2. The proportion p_i of the population living in each province (the word province is used in this section to denote the 10 Canadian provinces and the two territories) is also provided. The coefficients of variation (CV) of Table 2 were calculated from variances estimated with 5 random groups. Thus, one can consider that the sampling variances have a χ^2_4 distribution. Throughout this section, we assume that the provincial under-coverage estimates and their variances are independent.

Several estimators for provincial under-coverage are proposed by Royce (1992). Rivest (1995) proposed a composite estimator that shrinks the provincial under-coverage rate towards the national rate. It is given by:

$$r_i^c = \hat{\alpha} r_i + (1 - \hat{\alpha}) r_N,$$

where $r_N = \sum p_i r_i$ is the national under-coverage rate and the shrinking parameter $\hat{\alpha}$ is given by:

$$\hat{\alpha} = \frac{\sum p_i r_i^2 - r_N^2}{\sum p_i (1 - p_i) \sigma_i^2 + \sum p_i r_i^2 - r_N^2}.$$

This is the value of α that is optimal for loss functions for the estimation of provincial totals and of provincial shares of the population; see Royce (1992) and Rivest (1995) for details. One has $r_i^c = r_i + g_i(r)$, where

$$g_i(r) = - \frac{\sum p_i (1 - p_i) \sigma_i^2}{\sum p_i (1 - p_i) \sigma_i^2 + \sum p_i r_i^2 - r_N^2} (r_i - r_N).$$

A closed form expression for the conditional mean square error estimator can be calculated easily by noting that

$$\frac{\partial g_i(r)}{\partial r_i} = 2p_i(r_i - r_N)^2 \frac{\sum p_i (1 - p_i) \sigma_i^2}{\left[\sum p_i (1 - p_i) \sigma_i^2 + \sum p_i r_i^2 - r_N^2 \right]^2} - (1 - p_i)(1 - \hat{\alpha}).$$

The second partial derivative of $g_i(r)$ can also be calculated; it has the same sign as $r_i - r_N$. Thus positive skewness in the under-coverage rate, that is likely when estimating rare events such as being missed by the census, increases the conditional mean squared error in provinces where the under-coverage is above the national rate.

For 1991, $\hat{\alpha} = .874$ and the national under-coverage rate is $r_N = 2.872\%$. Table 2 gives the provincial composite under-coverage estimates, r_i^c together with their efficiencies $\text{eff}_{ic}^c = \sigma_{ii} / \text{mse}_c(r_i^c)$, where $\text{mse}_c(r_i^c)$ is calculated as defined in section 2, with the correction proposed in section 3.2 to account for estimated variances. The composite estimator is an improvement over the direct estimators in all cases except three, that correspond to the provinces with the most extreme under-coverage rates.

Table 2 also gives the empirical Bayes estimator r_i^B calculated with a location smoothing model. Under model (M), the true under-coverage rate θ_i is assumed to be distributed as a $N(\beta, \sigma_v^2)$. The parameter estimates are $\hat{\sigma}_v^2 = 1.45 \times 10^{-4}$ and $\hat{\beta}_w = 2.61\%$. Two efficiencies with respect to direct estimators are presented, eff_{ic}^B which is calculated with the conditional mean squared error estimator for r_i^B , including the adjustment of section 3.2 to account for estimated variances, and eff_{iPR}^B which is calculated with Prasad-Rao unconditional estimator. The large under-coverage rate in the N.W. Territories is responsible for the large estimate for $\hat{\sigma}_v^2$; this makes the empirical Bayes estimators r_i^B much closer to the direct estimators r_i than the composite estimators. Redoing the analysis without the N.W. Territories and Yukon changes the empirical Bayes estimates drastically.

Table 2
Two Estimators of Provincial Under-Coverage and Their Efficiencies

PROVINCE	p_i	r_i	CV	r_i^c	eff_{ic}^c	r_i^B	eff_{ic}^B	eff_{iPR}^B
Newfoundland	2.06	1.994	15.96	2.105	1.12	2.038	1.07	1.04
Prince Edward Island	0.47	0.931	30.00	1.176	0.65	1.025	0.93	1.03
Nova Scotia	3.26	1.889	20.05	2.013	1.11	1.959	1.09	1.06
New Brunswick	2.66	3.245	13.73	3.198	1.29	3.162	1.14	1.09
Québec	25.19	2.605	8.35	2.639	1.16	2.605	1.04	1.02
Ontario	37.24	3.641	8.46	3.544	0.87	3.572	1.02	1.04
Manitoba	3.96	1.86	20.83	1.987	1.10	1.936	1.09	1.06
Saskatchewan	3.58	1.798	18.87	1.933	1.04	1.863	1.06	1.05
Alberta	9.24	1.995	14.57	2.106	1.01	2.032	1.06	1.03
British Columbia	12.01	2.733	9.86	2.751	1.26	2.727	1.07	1.03
Yukon	0.10	3.83	15.99	3.709	1.27	3.56	1.03	1.17
N.W. Territories	0.22	5.439	11.28	5.116	0.96	4.813	0.49	1.18

In Table 2, the composite estimator performs better than the empirical Bayes estimator; it provides gains in conditional efficiency larger than 10% in 7 of 12 provinces. Three efficiencies are smaller than 1; the discussion in section 4.2 suggests that efficiencies less than 1 are unavoidable. The relatively poor precision of $\hat{\sigma}_{ii}$ (they are estimated using only 4 degrees of freedom), lowers the conditional efficiencies of the empirical Bayes estimators. It does not affect the composite estimator as much since it uses the same shrinking parameter for all provinces. The conditional efficiencies capture the poor performances of the r_i^c and r_i^B in the provinces with the most extreme under-coverage rates. This is missed by the Prasad Rao efficiencies. They highlight the gains that smoothing brings to the two territories where the under-coverage rates are highly variable. The Prasad Rao efficiencies are meaningful only if one accepts the hypothesis of provincial exchangeability underlying the smoothing model. This is doubtful since under-coverage tends to be higher in large urban provinces than in small rural areas.

5.2 Sub-Provincial Estimations

Dick (1995) considered the estimation of the adjustment factors for census under-coverage for age \times sex categories within each province for the 1991 census. The adjustment factor for a small area is defined as $F=1+$ (estimated under-coverage)/(census count). With four age categories, 0-19, 20-29, 30-44, 45+, and two sexes, there are 96 small areas. The explanatory variables are interactions between the indicator variables for the 12 provinces, the 4 age groups and the two sexes, and the proportions of renters (R) and of people that do not speak either official language (L) in the 96 small areas. In each one, the estimated variance was given by $\hat{\sigma}_{ii} = (\text{under-coverage variance}) / (\text{census count})^2$.

Dick (1995) regressed the log-variances on the census count to smooth the variance. He considered the exponentials of the predicted values for the log-variances ($\hat{\sigma}_{ii}$) as the known variances. This underestimates the variability.

Indeed, the average predicted variance $\hat{\sigma}_{ii}$ represents only 68% of the average unsmoothed variance. Multiplying $\hat{\sigma}_{ii}$ by $\exp(\hat{\sigma}_v^2/2) = 1.54$, where $\hat{\sigma}_v^2$ is the residual variance of the smoothing model, corrects this problem. Fitting Dick's (1995) model using the "unbiased" smoothed variance yields $\hat{\sigma}_v^2 = 0$. This is a degenerate situation where empirical Bayes estimators are equal to linear model predicted values. Note also the correlation between the variance residuals and the partial derivatives of g_i , calculated as if $\hat{\sigma}_v^2 > 0$, is 0.25. This suggest that (5) is violated. Using $\hat{\sigma}_{ii} \exp(\hat{\sigma}_v^2/2)$ in the calculation is likely to overestimate the precision the small area estimates. To illustrate the application of the conditional mean squared error estimator, these problems are ignored and the remainder of

this section assumes that the sampling variances σ_{ii} are known and equal to their smoothed values $\hat{\sigma}_{ii}$.

The model fitted by Dick (1995) has ten independent variables; the weighted least squares estimates and their standard errors, given by the square roots of the elements on the diagonal matrix of \hat{A}^{-1} , appear in Table 3. The conditional mean squared errors $\text{mse}_c^+(\hat{\mu}_i)$ for the 96 small areas can be calculated using (8). One had $\text{mse}_c^+(\hat{\mu}_i) = 0$ and $\text{mse}_c^+(\hat{\mu}_i) > \sigma_{ii}$ for respectively 51 and 15 small areas. The 15 small areas with large conditional mean squared errors need special attention: can the prediction model be improved for these areas? Two systematic features among the 15 corresponding residuals are noteworthy: there are 2 large positive residuals in the M/0-19 category and 2 large negative residuals in the F/45+ category. This suggests adding M/0-19 and F/45+ as independent variables. The additional column to the X matrix for M/0-19 contains 1's for the 12 small areas for males between 0 and 19 years old and 0 elsewhere; that for F/45+ is constructed in a similar way. Only F/45+ improves the fit; adding this explanatory variable gives the modified Dick model of Table 3. The absolute value of the t -statistic for F/45+ is 3; this is clearly significant.

It is interesting to compare the conditional mean squared errors obtained with the modified Dick model with those for Dick's model. Using the modified model decreases mse_c^+ in 26 small areas and increases it in 21; showing a slight improvement with the modified model.

The sub-provincial empirical Bayes adjustment factors can be aggregated at the provincial level. Provincial adjustment factors F_p are given by

$$\hat{F}_p = \frac{\sum_i C_i \hat{F}_i}{\sum_p C_i}$$

where C_i represents the census count for the i -th small area and \sum_p is the summation over the 8 small areas in province p . A mean squared error for the provincial adjustment factor, either conditional or unconditional, can be calculated using a mean product error matrix mpe as

$$\text{mse}(\hat{F}_p) = \frac{1}{\left(\sum_p C_i\right)^2} \sum_p \sum_p C_i C_j \text{mpe}(\hat{F}_i, \hat{F}_j).$$

Conditional mean squared errors are obtained by using formula (2) for mpe. Lahiri and Rao (1995) give a formula for the off-diagonal terms of the unconditional mean product error matrix whose diagonal is given by Prasad Rao (1990) mean squared errors.

Table 3
Two Linear Models for Small Area Correction Factors:
Dick ($p=11$) and Modified Dick ($p=12$). Parameter Estimates
are Given With Their Standard Errors in Parentheses

Category	Variable	Dick		modified Dick	
mean	intercept	1.0076	(0.0018)	1.0099	(0.0018)
Age* Sex Interaction	M / 20-29	0.0563	(0.0038)	0.0541	(0.0037)
	M / 30-44	0.0207	(0.0036)	0.0185	(0.0035)
	F / 20-20	0.0243	(0.0038)	0.02223	(0.0037)
	F / 45+	-	-	-0.0102	(0.0037)
Province* Renters Interaction	BC*R	0.0436	(0.0115)	0.0433	(0.0110)
	Ontario*R	0.0791	(0.0100)	0.0789	(0.0102)
	Québec*R	0.0253	(0.0097)	0.0259	(0.0090)
	N.-B*R	0.1039	(0.0194)	0.1032	(0.0186)
	Yukon*R	0.0633	(0.0179)	0.0634	(0.0175)
	NWT*R	0.0687	(0.0117)	0.0680	(0.0285)
Language*Sex*Age Interaction	L*F / 0-19	0.0802	(0.0293)	0.0680	(0.0285)
Variance		3.3681e-05	(2.45e-05)	2.21e-05	(2.30e-05)

Table 4

Direct (F_p) and Empirical Bayes (F_p^b) Estimates of the Provincial Correction Factors With Their Conditional (eff_{pc}) and Their Unconditional (eff_{pPR}) Efficiencies. A Conditional Efficiency is ∞ When the Conditional Mean Squared Error Estimator is Null

PROVINCE	F_p	F_p^b	eff_{pc}	eff_{pPR}
Newfoundland	1.0203	1.0176	6.49	2.94
Prince Edward Island	1.0094	1.0153	1.03	4.52
Nova Scotia	1.0193	1.0171	25.3	2.59
New Brunswick	1.0335	1.0367	0.67	1.11
Québec	1.0268	1.0262	1.12	0.93
Ontario	1.0378	1.0396	0.68	0.93
Manitoba	1.0190	1.0176	∞	2.46
Saskatchewan	1.0183	1.0166	∞	2.54
Alberta	1.0204	1.0187	7.37	1.98
British Columbia	1.0281	1.0293	1.09	1.03
Yukon	1.0396	1.0400	1.41	1.17
N.W. Territory	1.0575	1.0550	1.40	1.32

Direct and empirical Bayes aggregated estimates are presented in Table 4 with two efficiencies. The empirical Bayes estimates retain much of the interprovincial differences. This suggest that the explanatory variables of the smoothing model have captured most of the differences between the provincial under-coverage rates. A notable exception is Prince Edward Island's small correction factor which is not accounted for by the explanatory variables. This is the only province for which the two efficiencies differ substantially. The conditional efficiencies are more unstable than the Prasad Rao efficiencies. Except in Prince Edward Island, both tell similar stories: in New Brunswick, Quebec, Ontario, and British Columbia, the aggregated empirical Bayes estimates do not improve much on the direct estimators.

6. CONCLUSIONS

The estimator of the conditional mean squared error proposed in this paper has several interesting features. It can be implemented with any shrinking strategy. It is conditional on the realization of the smoothing model used to produce the small area characteristics; thus the conditional estimator has a large sampling variance. Simple modifications to the estimator are available to handle skewness in the data and estimated variances. In an empirical Bayes setting, it provides diagnostic information concerning the smoothing model. It can also be used as building blocks for estimators of the prediction variances when this variance has no closed form expression.

ACKNOWLEDGEMENTS

We are grateful to Peter Dick for providing the data set analyzed in section 5.2, and to Jon Rao for pointing out the instability of the conditional estimator under heavy smoothing. The financial contributions of the Fonds pour la formation des chercheurs et l'aide à la recherche du Québec and of the National Science and Engineering Research Council of Canada are gratefully acknowledged.

APPENDIX

Proof of Proposition 1

Let $\Sigma^{1/2}$ be a symmetric square root for Σ , such that $(\Sigma^{1/2})^2 = \Sigma$ and $z = \Sigma^{-1/2}(y - \mu)$. Note that z has a $N_n(0, I)$ distribution. In terms of the random vector $z, E\{(y - \mu)g(y)'\} = \Sigma^{1/2}E\{zg(\mu + \Sigma^{1/2}z)\}$. Now the conditional expectation of $z_i g_i(\mu + \Sigma^{1/2}z)$ given $(z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n)$ is equal to

$$\int_R \frac{z_i \exp(-z_i^2/2)}{\sqrt{2\pi}} g_j(\mu + \Sigma^{1/2} z) dz_i.$$

Integrating by parts shows that the above integral is equal to

$$\int_R \frac{\exp(-z_i^2/2)}{\sqrt{2\pi}} \frac{\partial g_j(\mu + \Sigma^{1/2} z)}{\partial z_i} dz_i.$$

Observe that

$$\frac{\partial g_j(\mu + \Sigma^{1/2} z)}{\partial z_i} = \sum_{k=1}^n \Sigma_{ki}^{1/2} g_j^k(\mu + \Sigma^{1/2} z).$$

Since $\Sigma^{1/2}$ is symmetric, $\Sigma_{ki}^{1/2} = \Sigma_{ik}^{1/2}$. Thus the above expression is the scalar product between $e_i' \Sigma^{1/2}$, the i -th row of $\Sigma^{1/2}$ (e_i represents a $n \times 1$ vector of 0's except for the i -th component which is 1), and $\nabla g(\mu + \Sigma^{1/2} z) e_j$, the j -th column of $\nabla g(y)$, evaluated at $y = \mu + \Sigma^{1/2} z$. We have

$$E\{z_i g_j(\mu + \Sigma^{1/2} z) \mid z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n\} = e_i' \Sigma^{1/2} E\{\nabla g(\mu + \Sigma^{1/2} z) e_j \mid z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n\}.$$

This equality also holds unconditionally, $E\{z_i g_j(\mu + \Sigma^{1/2} z)\} = e_i' \Sigma^{1/2} E\{\nabla g(\mu + \Sigma^{1/2} z) e_j\}$. In other words,

$$E\{zg(\mu + \Sigma^{1/2} z)\} = \Sigma^{1/2} E\{\nabla g(\mu + \Sigma^{1/2} z)\}.$$

This completes the proof.

Proof of Proposition 2

Let E_i denote the conditional expectation with respect to y_i , given $(y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n)$ and $h(y_i) = g_i(y)$, for fixed values of $(y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n)$. One has

$$E_i\{(y_i - \mu_i) h(y_i)\} = \int_R (t - \mu_i) h(t) f(t) dt.$$

To evaluate this expression, one can integrate by parts. Integrating $(t - \mu_i) \exp\{-(t - \mu_i)^2/(2\sigma_{ii})\}/(2\pi\sigma_{ii})^{1/2}$ in the above integrand yields

$$E_i\{(y_i - \mu_i) h(y_i)\} = \sigma_{ii} E_i\{h'(y_i)\} + \frac{\sigma_{ii}^{3/2} \rho_i}{2} \times \int_R h(t) \left\{ \frac{(t - \mu_i)^2}{\sigma_{ii}} - 1 \right\} \frac{\exp\{(t - \mu_i)^2/(2\sigma_{ii})\}}{(2\pi\sigma_{ii})^{1/2}} dt,$$

where $h'(t)$ is the derivative of $h(t)$. Repeated integrations by parts show that

$$\begin{aligned} & \int_R h(t) \frac{(t - \mu_i)^2}{\sigma_{ii}} \frac{\exp\{(t - \mu_i)^2/(2\sigma_{ii})\}}{(2\pi\sigma_{ii})^{1/2}} dt \\ &= \int_R \{h'(t)(t - \mu_i) + h(t)\} \frac{\exp\{(t - \mu_i)^2/(2\sigma_{ii})\}}{(2\pi\sigma_{ii})^{1/2}} dt \\ &= \int_R \{\sigma_{ii} h''(t) + h(t)\} \frac{\exp\{(t - \mu_i)^2/(2\sigma_{ii})\}}{(2\pi\sigma_{ii})^{1/2}} dt \end{aligned}$$

where $h''(t)$ is the second derivative of $h(t)$. This yields

$$E_i\{(y_i - \mu_i) h(y_i)\} = \sigma_{ii} E_i\{h'(y_i)\} + \frac{\sigma_{ii}^{3/2} \rho_i}{2} E_i\{h''(y_i)\} + o(\rho_i).$$

Taking, on both sides, the expectation with respect to the distribution of $(y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n)$ completes the proof.

Proof of Proposition 3

Let E_i denote the expectation taken with respect to the distribution of $\hat{\sigma}_{ii}$, given all the other random quantities $(y, \hat{\sigma}_{jj}, j \neq i)$. In this context one can write $(\partial g_i(y))/(\partial y_i) = h(\hat{\sigma}_{ii})$, where h is a function possibly depending on $(y, \hat{\sigma}_{jj}, j \neq i)$. A Taylor series expansion of h gives:

$$\begin{aligned} h(\hat{\sigma}_{ii}) &= h(\sigma_{ii}) + h'(\sigma_{ii})(\hat{\sigma}_{ii} - \sigma_{ii}) \\ &+ h''(\sigma_{ii}) \frac{(\hat{\sigma}_{ii} - \sigma_{ii})^2}{2} + O((\hat{\sigma}_{ii} - \sigma_{ii})^3). \end{aligned}$$

Since $(k-1)\hat{\sigma}_{ii}/\sigma_{ii}$ follows a χ_{k-1}^2 distribution, $E_i\{(\hat{\sigma}_{ii} - \sigma_{ii})^2\} = 2\sigma_{ii}^2/(k-1)$, and the centered moments of higher orders are $O(1/k^2)$. The above expansion reduces to,

$$\sigma_{ii} E_i\{\partial g_i(y)/\partial y_i\} = \sigma_{ii} h(\sigma_{ii}) + h''(\sigma_{ii}) \frac{\sigma_{ii}^3}{k-1} + O(1/k^2)$$

It is clear that the bias of $\hat{\sigma}_{ii} h(\hat{\sigma}_{ii})$ as an estimator of this expression is $O(1/k)$, provided that $h'(\sigma_{ii}) \neq 0$. One has, neglecting $O(1/k^2)$ terms,

$$\begin{aligned} & E_i \left\{ \hat{\sigma}_{ii} h \left(\frac{(k-1)\hat{\sigma}_{ii}}{k+1} \right) \right\} \\ & \approx \sigma_{ii} h(\sigma_{ii}) + h'(\sigma_{ii}) E_i \left\{ \hat{\sigma}_{ii} \left(\frac{(k-1)\hat{\sigma}_{ii}}{k+1} - \sigma_{ii} \right) \right\} \\ & + \frac{h''(\sigma_{ii})}{2} E_i \left\{ \hat{\sigma}_{ii} \left(\frac{(k-1)\hat{\sigma}_{ii}}{k+1} - \sigma_{ii} \right)^2 \right\} \end{aligned}$$

Elementary manipulations show that, in the above formula, the coefficient of $h'(\sigma_{ii})$ is null and

$$E_i \left\{ \hat{\sigma}_{ii} \left(\frac{(k-1)\hat{\sigma}_{ii}}{k+1} - \sigma_{ii} \right)^2 \right\} = 2 \frac{\sigma_{ii}^3}{k-1} + O(1/k^2).$$

This shows that

$$E_i \left\{ \hat{\sigma}_{ii} h \left(\frac{(k-1)\hat{\sigma}_{ii}}{k+1} \right) \right\} = \sigma_{ii} E_i \{ \partial g_i(y) / \partial y_i \} + O(1/k^2).$$

The proof is completed by noting that this equality holds for the unconditional expectation, taken with respect to the joint distribution of $(y, \hat{\sigma}_{ii}, i = 1, \dots, n)$.

Proof of Proposition 4

The mean squared error of the posterior variance as an estimator of the conditional mean squared error has only a bias term,

$$\left(\frac{\sigma^2}{\sigma^2 + \sigma_v^2} \right)^4 \{ (\mu_i - x_i' \beta)^2 - \sigma_v^2 \}^2,$$

while the mean squared error of $\text{mse}_c(\hat{\mu}_i)$ has only a variance component which is given by

$$\begin{aligned} & \left(\frac{\sigma_2}{\sigma^2 + \sigma_v^2} \right)^4 \text{Var}_S \{ (y_i - x_i' \beta)^2 \} \\ &= \left(\frac{\sigma^2}{\sigma^2 + \sigma_v^2} \right)^4 \{ 2\sigma_{ii}^2 + 4(\mu_i - x_i' \beta)^2 \sigma^2 \}. \end{aligned}$$

The efficiency reported in Proposition 4 can be evaluated as the ratio of the 2 average mean squared errors defined above. It is given by,

$$\frac{2\sigma^4 + 4\sigma^2 \sum (\mu_i - x_i' \beta)^2 / n}{\sum \{ (\mu_i - x_i' \beta)^2 - \sigma_v^2 \} / n}.$$

Taking expectations of the numerator and of the denominator with respect to model (6) yields the result.

REFERENCES

- BARNDORFF-NIELSEN, O.E., and COX, D.R. (1989). *Asymptotic Techniques for Use in Statistics*. New York: Chapman and Hall.
- BELMONTE, E. (1998). Estimation dans les petites régions: une nouvelle dérivation de l'erreur quadratique moyenne de Prasad-Rao. 1998 *Proceedings of the Survey Methods Section, Statistical Society of Canada*, 165-170.
- BELMONTE, E. (1999). *L'estimation dans les petites régions: Survol des méthodes de Bayes et présentation d'un estimateur conditionnel de l'EQM*. Mémoire de maîtrise. Département de mathématiques et de statistique, Université Laval.
- BILODEAU, M., and SRIVASTAVA, M.S. (1988). Estimation of the MSE matrix of the Stein estimator. *Canadian Journal of Statistics*, 16, 153-159.
- BOOTH, J.G., and HOBERT, J.P. (1998). Standard errors of predictions in generalized linear mixed models. *Journal of the American Statistical Association*, 93, 262-272.
- BURGESS, R.D. (1988). Evaluation of the reverse record check estimates of under-coverage in the Canadian Census of Population. *Survey Methodology*, 14, 137-156.
- CRESSIE, N. (1992). REML estimation in empirical Bayes smoothing of census undercount. *Survey Methodology*, 18, 75-94.
- DICK, P. (1995). Modeling net under-coverage in the 1991 Canadian Census. *Survey Methodology*, 21, 44-55.
- FAY, R.E., and HERRIOT, R.A. (1979). Estimates of income for small places: An application of James Stein procedure to census data. *Journal of the American Statistical Association*, 74, 269-277.
- GHOSH, M., and RAO, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-93.
- HOGAN, H. (1992). The 1990 Post-Enumeration Survey: an overview. *The American Statistician*, 46, 261-269.
- KACKAR, R.N., and HARVILLE, D.A. (1984). Approximations for standard errors of estimators for fixed and random effects in mixed models. *Journal of the American Statistical Association*, 79, 853-862.
- KOTT, P.S. (1989). Robust small domain estimation using random effect modeling. *Survey Methodology*, 15, 3-12.
- LAHIRI, P., and RAO, J.N.K. (1995). Robust estimation of mean squared error of small area estimators. *Journal of the American Statistical Association*, 90, 758-766.
- MARITZ, J.S., and LWIN, T. (1989). *Empirical Bayes Methods*. (Second Edition), London: Chapman and Hall.
- PRASAD, N.G.N., and RAO, J. N. K. (1990). The estimation of mean squared errors of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.
- PRASAD, N.G.N., and RAO, J.N.K. (1999). On robust small area estimation using a simple random effect model. *Survey Methodology*, 25, 163-171.
- PURCELL, N.J., and KISH, L. (1979). Estimation for small domains. *Biometrics*, 35, 365-384.
- RAO, C.R., and SHINOZAKI, N. (1978). Precision of individuals estimators in simultaneous estimation of parameters. *Biometrika*, 65, 23-30.
- REID, N. (1991). Approximations and asymptotics. In *Statistical Theory and Modeling. In Honor of Sir David Cox, FRS*, (eds. D.V. Hinkley, N. Reid and E.J. Snell), 287-305.
- RIVEST, L.P. (1995). A composite estimator for provincial under-coverage in the Canadian census. 1995 *Proceedings of the Survey Methods Section. Statistical Society of Canada*, 33-38.
- ROBERT, C. (1992). *L'Analyse Statistique Bayésienne*. Paris: Economica.
- ROYCE, D. (1992). A comparison of some estimators for a set of population totals. *Survey Methodology*, 18, 109-125.
- SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

- SCOTT, A., and SMITH T.M.F. (1971). Interval estimates for linear combinations of means. *Applied Statistics*, 20, 276-285.
- SINGH, M.P., GAMBINO, J., and MANTEL, H.J. (1994). Issues and strategy for small area data. *Survey Methodology*, 20, 1-22.
- SINGH, A.C., STUKEL, D.M., and PFEFFERMANN, D. (1998). Bayesian versus frequentist measures of error in small area estimation. *Journal of the Royal Statistical Society, Series B*, 60, 377-396.
- STEIN, C. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9, 1135-1151.
- WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

Cold Deck and Ratio Imputation

JUN SHAO¹

ABSTRACT

Imputation is a common procedure to compensate for nonresponse in survey problems. Using auxiliary data, imputation may produce estimators that are more efficient than the one constructed by ignoring nonrespondents and re-weighting. We study and compare the mean squared errors of survey estimators based on data imputed using three different imputation techniques: the commonly used ratio imputation method and two cold deck imputation methods that are frequently adopted in economic area surveys conducted by the U.S. Census Bureau and the U.S. Bureau of Labor Statistics. A cold deck method imputes a nonrespondent of an item by reported values from anything other than reported values for the same item in the current data set (e.g., values from a covariate and/or from a previous survey). Although sometimes a cold deck imputation method makes use of more auxiliary data than the other imputation methods, it is not always better in terms of the mean squared errors of the resulting survey estimators. In a simple case we compare explicitly the mean squared errors and discuss situations under which one method is better than the other two. In general cases we propose to compare mean squared errors empirically based on some consistent estimates of mean squared errors. Estimation of mean squared errors of survey estimators in the presence of imputed data is itself an important problem in surveys. A numerical example related to the Transportation Annual Survey is presented for illustration.

KEY WORDS: Complex survey; Mean squared error; Nonresponse; Simple random sample; Variance estimation.

1. INTRODUCTION

Imputation is one of the most common procedures to compensate for nonresponse in survey problems. In addition to many practical reasons for imputation, imputation using auxiliary data may produce estimators that are more efficient than the one constructed by ignoring nonrespondents and re-weighting. Suppose that we have a sample s selected from a finite population \mathcal{P} consisting of some units represented by $i = 1, \dots, M$, and that we observe $\{y_i, i \in r\}$ (respondents), $r \subset s$. Suppose also that we have auxiliary data x_i 's observed for all $i \in s$ and $x_i > 0$. The commonly used ratio imputation method (see, for example, Kalton and Kasprzyk 1986) imputes nonrespondents as follows. First, we create K imputation cells $\mathcal{P}_1, \mathcal{P}_2 \cup \dots \cup \mathcal{P}_K = \mathcal{P}$, according to a categorical auxiliary variable (which is observed for every $i \in s$ and is typically different from x) such that for every k , the following model is assumed to hold:

$$y_i = \beta_k x_i + x_i^{1/2} e_i, \quad i \in \mathcal{P}_k \quad (1)$$

$$P(a_i = 1 | y_i, x_i) = P(a_i = 1 | x_i),$$

where β_k is an unknown parameter, e_i is independent of x_i with $E(e_i) = 0$ and unknown $V(e_i) = \sigma_k^2 > 0$, a_i is the indicator of whether y_i is a respondent, and (a_i, x_i) 's are independent. Then, within imputation cell k , a nonrespondent y_i is imputed by $\hat{\beta}_k x_i$, where

$$\hat{\beta}_k = \sum_{i \in r_k} w_i y_i / \sum_{i \in r_k} w_i x_i \quad (2)$$

is the best linear unbiased estimator of β_k under model (1), r_k is r restricted to the k -th imputation cell, and w_i is the survey weight associated with the i -th sampled unit. Note that model (1) consists of a regression model between y_i and x_i (with no intercept and with error variance proportional to x_i) and a response model which assumes that the response mechanism is independent of y_i 's, given x_i 's. This response mechanism is termed as missing at random by Rubin (1976) or unconfounded response mechanism by Lee, Rancourt and Särndal (1994). Based on the imputed data set, the Horvitz-Thompson (HT) estimator of Y , the population total of y_i 's, is

$$\hat{Y}_R = \sum_k \left(\sum_{i \in r_k} w_i y_i + \sum_{i \in s_k - r_k} w_i \hat{\beta}_k x_i \right), \quad (3)$$

where s_k is s restricted to the k -th imputation cell. The HT estimator of Y obtained by ignoring nonrespondents and re-weighting within each imputation cell is

$$\hat{Y}_w = \sum_k \sum_{i \in r_k} \tilde{w}_{ik} y_i, \quad \tilde{w}_{ik} = w_i \left(\sum_{i \in s_k} w_i / \sum_{i \in r_k} w_i \right). \quad (4)$$

It can be seen that if $x_i \equiv 1$, then the estimators in (3) and (4) are the same. Both estimators are unbiased if model (1) holds. (Throughout this paper, the bias and variance are with respect to model (1) and repeated sampling, unless otherwise specified.) Under model (1), however, \hat{Y}_R is more efficient than \hat{Y}_w if the size of r is substantially smaller than the size of s . Even if the regression model in (1) does not hold, \hat{Y}_R may still be more efficient than \hat{Y}_w in terms of their mean squared errors with respect to repeated sampling (Cochran 1977, Chapter 6) when the response

¹ Jun Shao, Department of Statistics, University of Wisconsin, Madison, WI 53706 U.S.A. E-mail: shao@stat.wisc.edu.

probability is a constant in any given imputation cell (which ensures that \hat{Y}_R and \hat{Y}_W are approximately unbiased with respect to repeated sampling).

The purpose of this note is to compare the efficiency of \hat{Y}_R with other estimators of Y based on data with nonrespondents imputed by using a method called cold deck. A cold deck method imputes a nonrespondent of y -variable by reported values from anything other than y -values (e.g., values from a covariate and/or from a previous survey). Cold deck imputation is opposite to hot deck imputation in which a nonrespondent is imputed by a respondent from the same variable in the current survey. The ratio imputation method uses both reported y -values and auxiliary data and is sometimes called a "warm deck" method. The simplest cold deck imputes a nonrespondent y_i , $i \in s - r$, by x_i and the resulting HT estimator of Y is

$$\hat{Y}_C = \sum_{i \in r} w_i y_i + \sum_{i \in s - r} w_i x_i. \quad (5)$$

The use of this simple cold deck is motivated by the fact that under model (1), β_k 's are close to 1 in many survey problems, especially when x_i 's are y -values from a previous survey. When some β_k 's are not equal to 1, \hat{Y}_C in (5) has a bias which does not vanish even if $s = \mathcal{P}$ (i.e., the sample is a census). However, having a small bias may be paid off by lowering the variance so that the overall mean squared error $\text{mse}(\hat{Y}_C) = E(\hat{Y}_C - Y)^2$ may still be smaller than the mean squared error $\text{mse}(\hat{Y}_R) = E(\hat{Y}_R - Y)^2 = V(\hat{Y}_R - Y)$, where E and V denote the expectation and variance under model (1) and repeated sampling. More details can be found in section 2. The simple cold deck may be improved by another cold deck method, the cold deck-ratio method, which imputes a nonrespondent y_i by $x_i \tilde{y}_i / \tilde{x}_i$, where \tilde{y}_i and \tilde{x}_i are reported values from a previous survey. The corresponding HT estimator of Y is

$$\hat{Y}_{C-R} = \sum_{i \in r} w_i y_i + \sum_{i \in s - r} w_i x_i \tilde{y}_i / \tilde{x}_i. \quad (6)$$

The estimator in (6) is unbiased if model (1) holds for \tilde{y}_i and \tilde{x}_i (i.e., $\tilde{y}_i = \beta_k \tilde{x}_i + \tilde{x}_i^{1/2} \tilde{e}_i$) with the same β_k as the one for y_i and x_i . These two cold deck methods are widely used in economic area surveys conducted by the U.S. Census Bureau (King and Kornbau 1994) and the U.S. Bureau of Labor Statistics (Butani, Harter and Wolter 1998). Applying cold deck imputation methods does not require knowing the imputation cells, although model (1) is assumed to ensure the unbiasedness of \hat{Y}_C and \hat{Y}_{C-R} .

Although the cold deck-ratio method makes use of more auxiliary data, it is not always better than the simple cold deck or the ratio imputation method. In section 2 we compare explicitly the mean squared errors of \hat{Y}_R , \hat{Y}_C and \hat{Y}_{C-R} in a special case where the sample s is a simple random sample (SRS) and the response probability is a constant. Situations under which one method is better than the others are discussed. If the sampling design or the response mechanism is complex, then it is not easy to

compare the mean squared errors explicitly. One may, however, estimate the mean squared errors of \hat{Y}_R , \hat{Y}_C and \hat{Y}_{C-R} and make an empirical comparison. Variance or mean squared error estimation is itself an important problem, since it is common to report variance or mean squared error estimates along with the estimated totals. These are discussed in section 3.

Our results can also be applied to the problem related to two-phase sampling or double sampling, which is often employed when it is cheap to take a large sample $\{x_i, i \in s\}$ and expensive to obtain y -values so that a subsample $\{y_i, i \in r\}$ is taken in the second-phase, $r \subset s$.

A numerical example is discussed in section 4 using data from the Transportation Annual Survey conducted by the U.S. Census Bureau.

2. SRS WITH UNIFORM RESPONSE

To illustrate the idea, we start with the simplest case where s is an SRS (without replacement from \mathcal{P} but the sampling fraction is negligible); there is only one imputation cell so that we can drop the subscript k for imputation cell; and the response probability is a constant $p > 0$ (uniform response mechanism).

In this case $w_i = N/n$, where n is the size of the sample s and N is the size of the population \mathcal{P} . Since $n/N \approx 0$ is assumed,

$$\text{mse}(\hat{Y}_R) \approx \frac{N^2}{n} \left(\frac{\sigma^2 \mu_x}{p} + \beta^2 v_x \right) \quad (7)$$

for large n , where $\mu_x = E(x_i)$ and $v_x = V(x_i)$ and, throughout the paper, $A \approx B$ means that A is equal to B up to a term which is relatively negligible compared to A and B as all sample sizes in imputation cells increase to infinity. A more detailed derivation of result (7) is given in the Appendix. For \hat{Y}_W in (4), it is easy to see that $\tilde{w}_i = N/r$, where r is the size of r , and \hat{Y}_W is unbiased. Then

$$\text{mse}(\hat{Y}_W) = V(\hat{Y}_W - Y) \approx V(\hat{Y}_W) = \frac{N^2}{n} \left(\frac{\sigma^2 \mu_x}{p} + \frac{\beta^2 v_x}{p} \right).$$

Hence \hat{Y}_R is more efficient than \hat{Y}_W unless $p = 1$ and $\beta^2 v_x = 0$. The gain in using \hat{Y}_R is proportional to β^2 and v_x , both are measures of usefulness of the auxiliary variable x in explaining y through model (1).

For the simple cold deck,

$$\hat{Y}_C = \frac{N}{n} \left(\sum_{i \in r} y_i + \sum_{i \in s - r} x_i \right) = \frac{N}{n} \left(\sum_{i \in r} x_i^{1/2} e_i + \beta \sum_{i \in r} x_i + \sum_{i \in s - r} x_i \right),$$

where e_i 's are defined in (1). Consequently,

$$V(\hat{Y}_C) = \frac{N^2}{n} \left\{ \sigma^2 p \mu_x + (\beta^2 p + 1 - p) v_x + (\beta - 1)^2 p (1 - p) \mu_x^2 \right\} \quad (8)$$

(see the Appendix). The bias of \hat{Y}_C is

$$E(\hat{Y}_C - Y) = N\mu_x(1-p)(1-\beta)$$

and, hence,

$$\begin{aligned} \text{mse}(\hat{Y}_C) &= V(\hat{Y}_C - Y) + [E(\hat{Y}_C - Y)]^2 \\ &\approx V(\hat{Y}_C) + [E(\hat{Y}_C - Y)]^2 \\ &= \frac{N^2}{n} \left\{ \sigma^2 p \mu_x + (\beta^2 p + 1 - p) v_x \right. \\ &\quad \left. + (\beta - 1)^2 (1 - p) [p + n(1 - p)] \mu_x^2 \right\}. \end{aligned} \quad (9)$$

Comparing (7) and (9), we obtain the following conclusions.

1. When $p = 1$ (no nonresponse), $\text{mse}(\hat{Y}_C) = \text{mse}(\hat{Y}_R)$.
2. When $p < 1$ and $\beta = 1$ (y and x have the same mean), $\text{mse}(\hat{Y}_C) < \text{mse}(\hat{Y}_R)$.
3. When $p < 1$ and $\beta \neq 1$, $\text{mse}(\hat{Y}_C) \leq \text{mse}(\hat{Y}_R)$ if and only if

$$(\beta - 1)^2 [p + n(1 - p)] \mu_x + (1 - \beta^2) v_x / \mu_x - \sigma^2 (p + 1) / p \leq 0. \quad (10)$$

Assume that $\mu_x > 0$. In most economic surveys, the relative variance v_x / μ_x^2 is smaller than $p + n(1 - p)$. Hence the left hand side of (10) is a quadratic function of β with a positive coefficient in the β^2 term and, therefore, the simple cold deck is better when β is in the interval with limits

$$\frac{[p + n(1 - p)] \mu_x \pm \sqrt{v_x^2 / \mu_x^2 + \{[p + n(1 - p)] \mu_x - v_x / \mu_x\} \sigma^2 (p + 1) / p}}{[p + n(1 - p)] \mu_x - v_x / \mu_x}.$$

This interval contains 1 since (10) holds if $\beta = 1$. Note that $[p + n(1 - p)] \mu_x$ increases to infinity as n increases to infinity. Hence the interval of β 's for which the simple cold deck is better shrinks to a single point ($\beta = 1$) as $n \rightarrow \infty$.

We now consider the cold deck-ratio. Assume that $\tilde{y}_i = \beta \tilde{x}_i + \tilde{x}_i^{1/2} \tilde{e}_i$, $E(\tilde{e}_i) = 0$, $V(\tilde{e}_i) = \sigma^2$, and that \tilde{e}_i , e_i , and (x_i, \tilde{x}_i) are mutually independent. Let $z_i = x_i \tilde{y}_i / \tilde{x}_i$ and $\epsilon_i = y_i - z_i = x_i^{1/2} e_i - \tilde{e}_i x_i / \tilde{x}_i^{1/2}$. Then $E(\hat{Y}_{C-R} - Y) = 0$ and

$$\text{mse}(\hat{Y}_{C-R}) = \frac{N^2}{n} \left\{ \sigma^2 p \mu_x + \beta^2 v_x + \sigma^2 (1 - p) \gamma_x \right\}, \quad (11)$$

where $\gamma_x = E(x_i^2 / \tilde{x}_i)$ (see the Appendix). By (7) and (11),

$$\text{mse}(\hat{Y}_R) - \text{mse}(\hat{Y}_{C-R}) = \frac{N^2 \sigma^2 (1 - p)}{n} \left\{ \left(\frac{1}{p} + 1 \right) \mu_x - \gamma_x \right\} \quad (12)$$

and, hence, the cold deck-ratio is better than the ratio imputation method if and only if $1/p + 1 \geq \gamma_x / \mu_x$. Note that

$\gamma_x \geq \mu_x$ and γ_x is close to μ_x if x_i and \tilde{x}_i are highly and positively related, in which case cold deck-ratio imputation can be much better than ratio imputation.

The comparison between the simple cold deck and the cold deck-ratio is the same as that between the simple cold deck and the ratio imputation method. One only needs to replace $(p + 1)/p$ in the third term of the left hand side of (10) by γ_x / μ_x .

The parameters β , σ , μ_x , v_x and γ_x have to be estimated in order to compare the efficiencies of \hat{Y}_R , \hat{Y}_C and \hat{Y}_{C-R} . Instead, we can directly compare estimated mean squared errors of \hat{Y}_R , \hat{Y}_C and \hat{Y}_{C-R} . This is discussed next.

3. STRATIFIED SAMPLING WITH UNCONFOUNDED RESPONSE

We consider the following stratified sampling design adopted by many U.S. government survey agencies: the finite population \mathcal{P} is stratified into H strata with N_h units in the h -th stratum; $n_h \geq 2$ units are selected without replacement from stratum h , according to some probability sampling plan; and the units are selected independently across the strata.

The survey weights w_i 's are constructed so that if all y_i 's are observed, the HT estimator $\sum_{i \in s} w_i y_i$ is unbiased for Y under repeated sampling.

We assume model (1). The response probability is no longer a constant, but independent of the y -value. For the cold deck-ratio, we also assume that within the k -th imputation cell, $\tilde{y}_i = \beta_k \tilde{x}_i + \tilde{x}_i^{1/2} \tilde{e}_i$, $E(\tilde{e}_i) = 0$, $V(\tilde{e}_i) = \sigma_k^2$ and e_i , \tilde{e}_i , (x_i, \tilde{x}_i) are mutually independent.

Explicit results for the mean squared errors such as (7), (9) and (11) are not easy to obtain. We may, however, make empirical comparisons of the efficiencies of \hat{Y}_R , \hat{Y}_C and \hat{Y}_{C-R} , based on their estimated mean squared errors. Estimation of the mean squared errors of \hat{Y}_R , \hat{Y}_C and \hat{Y}_{C-R} , is in fact an important part of the sampling theory. It is well known that for imputed data sets, the naive method that applies the standard variance estimation formulas by treating imputed nonrespondents as observed data leads to underestimation. When no correct method (for estimating the mean squared error) is available, the naive method is used in many survey agencies.

We now derive estimators for $V(\hat{Y})$ or $\text{mse}(\hat{Y})$ that are correct under model (1), where \hat{Y} denotes \hat{Y}_R , \hat{Y}_C or \hat{Y}_{C-R} .

Let E_m and V_m be the expectation and variance with respect to model (1) and let E_s and V_s be the expectation and variance with respect to repeated sampling (conditional on the model and response). Then

$$V(\hat{Y} - Y) = E_m[V_s(\hat{Y})] + V_m[E_s(\hat{Y}) - Y]. \quad (13)$$

We first consider $E_m[V_s(\hat{Y})]$, the first variance component in (13). It suffices to obtain an estimator of $V_s(\hat{Y})$, conditional on $\{y_i, x_i, a_i, i \in \mathcal{P}\}$ (and $\{\tilde{y}_i, \tilde{x}_i, i \in \mathcal{P}\}$ for cold deck-ratio), where a_i is the response indicator for y_i .

The estimation of $V_s(\hat{Y}_C)$ and $V_s(\hat{Y}_{C-R})$ is simple (which is an advantage of using a cold deck method). Let

$$v_1 = \sum_h \left(1 - \frac{n_h}{N_h} \right) \frac{n_h}{n_h - 1} \sum_{i \in s(h)} \left(w_i t_i - \frac{1}{n_h} \sum_{i \in s(h)} w_i t_i \right)^2 \quad (14)$$

be the standard variance estimator for $\sum_{i \in s} w_i t_i$ when $\{t_i, i \in s\}$ is treated as an observed sample (from $\{t_i, i \in \mathcal{P}\}$), where $s(h)$ is s restricted to stratum h . Then $V_s(\hat{Y}_C)$ can be estimated by using (14) with $t_i = a_i y_i + (1 - a_i)x_i$ and $V_s(\hat{Y}_{C-R})$ can be estimated by using (14) with $t_i = a_i y_i + (1 - a_i)x_i \tilde{y}_i / \tilde{x}_i$.

The estimation of $V_s(\hat{Y}_R)$ is slightly more complicated but similar. Assume that in each imputation cell, the number of sampled units is large and the response probabilities are bounded away from 0. Note that

$$\begin{aligned} \hat{Y}_R &= \sum_k \left[\left(\sum_{i \in s_k} w_i x_i / \sum_{i \in r_k} w_i x_i \right) \right. \\ &\quad \times \sum_{i \in r_k} w_i (y_i - \beta_k x_i) + \beta_k \sum_{i \in s_k} w_i x_i \left. \right] \\ &\approx \sum_k \left[\zeta_k \sum_{i \in s_k} w_i a_i (y_i - \beta_k x_i) + \beta_k \sum_{i \in s_k} w_i x_i \right] \\ &= \sum_{i \in s} w_i [\zeta_i a_i (y_i - \beta_i x_i) + \beta_i x_i], \end{aligned}$$

where $\zeta_k = E(\sum_{i \in s_k} w_i x_i) / E(\sum_{i \in r_k} w_i x_i)$ and $\zeta_i = \zeta_k$ and $\beta_i = \beta_k$ for $i \in s_k$. After estimating β_k by $\hat{\beta}_k$ and ζ_k by $\hat{\zeta}_k = \sum_{i \in s_k} w_i x_i / \sum_{i \in r_k} w_i x_i$, we estimate $V_s(\hat{Y}_R)$ by using (14) with $t_i = \hat{\zeta}_i a_i (y_i - \hat{\beta}_i x_i) + \hat{\beta}_i x_i$, where $\hat{\zeta}_i = \hat{\zeta}_k$ and $\hat{\beta}_i = \hat{\beta}_k$ for $i \in s_k$.

Before we discuss the estimation of $V_m[E_s(\hat{Y}) - Y]$, the second variance component in (13), it should be noted that $V_m[E_s(\hat{Y}) - Y] / E_m[V_s(\hat{Y})] = O(n/N)$. This is because the variance of $E_s(\hat{Y}) - Y$ (if it is nonzero) is typically of the order N , whereas the order of $V_s(\hat{Y})$ is typically N^2/n and thus the order of $E_m[V_s(\hat{Y})]$ is N^2/n under some regularity conditions. Hence, in theory, it is not necessary to estimate $V_m[E_s(\hat{Y}) - Y]$ if the sampling fraction n/N is negligible. However, the constant in $O(n/N)$ is unknown and, hence, one may still want to estimate $V_m[E_s(\hat{Y}) - Y]$ in applications even when n/N is small.

We now consider the estimation of the second variance component in (13). For \hat{Y}_C ,

$$\begin{aligned} E_s(\hat{Y}_C) - Y &= \sum_{i \in \mathcal{P}} [a_i y_i + (1 - a_i)x_i] - \sum_{i \in \mathcal{P}} y_i = \\ &\quad - \sum_{i \in \mathcal{P}} (1 - a_i)(y_i - x_i). \end{aligned}$$

Then, under model (1),

$$V_m[E_s(\hat{Y}_C) - Y] =$$

$$E_m \left[\sum_k \sigma_k^2 \sum_{i \in \mathcal{P}_k} (1 - a_i)x_i \right] + V_m \left[\sum_{i \in \mathcal{P}} (1 - a_i)(\beta_i - 1)x_i \right].$$

If we estimate σ_k^2 by

$$\hat{\sigma}_k^2 = \sum_{i \in s_k} a_i w_i (y_i - \hat{\beta}_k x_i)^2 / \sum_{i \in s_k} a_i w_i x_i,$$

then an estimator of $V_m[E_s(\hat{Y}_C) - Y]$ is

$$\begin{aligned} v_{2C} &= \sum_k \hat{\sigma}_k^2 \sum_{i \in s_k} (1 - a_i) w_i x_i + \\ &\quad \sum_h \frac{N_h}{n_h - 1} \sum_{i \in s(h)} \left(u_i - \frac{1}{n_h} \sum_{i \in s(h)} u_i \right)^2, \end{aligned} \quad (15)$$

where $u_i = (1 - a_i)(\hat{\beta}_i - 1)x_i$ and $\hat{\beta}_i = \hat{\beta}_k$ for $i \in s_k$.

For \hat{Y}_{C-R} ,

$$E_s(\hat{Y}_{C-R}) - Y = - \sum_{i \in \mathcal{P}} (1 - a_i)(y_i - x_i \tilde{y}_i / \tilde{x}_i)$$

and

$$\begin{aligned} V_m[E_s(\hat{Y}_{C-R}) - Y] &= \\ E_m \left[\sum_k \sigma_k^2 \sum_{i \in \mathcal{P}_k} (1 - a_i)x_i + \sum_k \hat{\sigma}_k^2 \sum_{i \in \mathcal{P}_k} (1 - a_i)x_i^2 / \tilde{x}_i \right]. \end{aligned}$$

Hence $V_m[E_s(\hat{Y}_{C-R}) - Y]$ can be estimated by

$$v_{2C-R} = \sum_k \left[\hat{\sigma}_k^2 \sum_{i \in s_k} (1 - a_i) w_i x_i + \hat{\sigma}_k^2 \sum_{i \in s_k} (1 - a_i) w_i x_i^2 / \tilde{x}_i \right], \quad (16)$$

where

$$\hat{\sigma}_k^2 = \sum_{i \in s_k} w_i (\tilde{y}_i - \hat{\beta}_k \tilde{x}_i)^2 / \sum_{i \in s_k} w_i \tilde{x}_i$$

and

$$\hat{\beta}_k = \sum_{i \in s_k} w_i \tilde{y}_i / \sum_{i \in s_k} w_i \tilde{x}_i.$$

For \hat{Y}_R ,

$$E_s(\hat{Y}_R) - Y \approx \sum_k \left[\left(\sum_{i \in \mathcal{P}_k} x_i / \sum_{i \in \mathcal{P}_k} a_i x_i \right) \sum_{i \in \mathcal{P}_k} a_i y_i - \sum_{i \in \mathcal{P}_k} y_i \right]$$

and from Taylor's expansion,

$$V_m[E_s(\hat{Y}_R) - Y] \approx$$

$$E_m \left[\sum_k \sigma_k^2 \left[\sum_{i \in \mathcal{P}_k} x_i \sum_{i \in \mathcal{P}_k} (1 - a_i)x_i \right] / \sum_{i \in \mathcal{P}_k} a_i x_i \right].$$

It can be estimated by

$$v_{2R} = \sum_k \hat{\sigma}_k^2 \left[\sum_{i \in s_k} w_i x_i \sum_{i \in s_k} (1 - a_i) w_i x_i \right] / \sum_{i \in s_k} a_i w_i x_i. \quad (17)$$

Finally, \hat{Y}_R and \hat{Y}_{C-R} are unbiased but \hat{Y}_C has a bias

$$\sum_k (1 - \beta_k) E_m \left[\sum_{i \in \mathcal{P}_k} (1 - a_i)x_i \right],$$

which can be estimated by

$$\sum_k (1 - \hat{\beta}_k) \sum_{i \in s_k} (1 - a_i) w_i x_i.$$

Thus, we obtain the following estimated mean squared errors: $\text{mse}(\hat{Y}_R)$ can be estimated by

$$\widehat{\text{mse}}(\hat{Y}_R) = v_{1R} + v_{2R},$$

where v_{1R} is obtained using (14) with $t_i = \hat{\zeta}_i a_i (y_i - \hat{\beta}_i x_i) + \hat{\beta}_i x_i$, $\hat{\zeta}_i = \hat{\zeta}_k$ and $\hat{\beta}_i = \hat{\beta}_k$ for $i \in s_k$, and v_{2R} is given by (17); $\text{mse}(\hat{Y}_C)$ by

$$\widehat{\text{mse}}(\hat{Y}_C) = v_{1C} + v_{2C} + \left[\sum_k (1 - \hat{\beta}_k) \sum_{i \in s_k - r_k} w_i x_i \right]^2,$$

where v_{1C} is obtained by using (14) with $t_i = a_i y_i + (1 - a_i) x_i$ and v_{2C} is given by (15); and $\text{mse}(\hat{Y}_{C-R})$ can be estimated by

$$\widehat{\text{mse}}(\hat{Y}_{C-R}) = v_{1C-R} + v_{2C-R},$$

where v_{1C-R} is obtained by using (14) with $t_i = a_i y_i + (1 - a_i) x_i \tilde{y}_i / \tilde{x}_i$ and v_{2C-R} is given by (16).

Under model (1) and the asymptotic settings in Krewski and Rao (1981), Rao and Shao (1992) or Valliant (1993), the derived mean squared error estimators are asymptotically unbiased and consistent as all sample sizes in imputation cell increase to infinity.

For cold deck or cold deck-ratio imputation, the first term (v_{1C} or v_{1C-R}) in the estimated mean squared error is the same as the one obtained by applying a standard formula (such as (14)) and treating imputed nonrespondents as observed data. For ratio imputation, applying (14) and treating imputed nonrespondents as observed data produces the following estimator of $\text{mse}(\hat{Y}_R)$:

$$\tilde{v}_{1R} = \sum_h \left(1 - \frac{n_h}{N_h} \right) \frac{n_h}{n_h - 1} \sum_{i \in s(h)} \left(w_i z_i - \frac{1}{n_h} \sum_{i \in s(h)} w_i z_i \right)^2 \quad (18)$$

with $z_i = a_i y_i + (1 - a_i) \hat{\beta}_i x_i$, which is different from the first term v_{1R} in our estimator $\widehat{\text{mse}}(\hat{Y}_R)$ and, hence, is not asymptotically valid even if n/N is negligible.

4. AN EXAMPLE

We consider an example using a data set from the Transportation Annual Survey (TAS) conducted by the U.S. Census Bureau.

The TAS is a survey of firms with one or more establishments that are primarily engaged in providing commercial motor freight transportation or public warehousing services in U.S. A stratified simple random sample is selected without replacement from employers contained in the Census Bureau's Standard Statistical Establishment List.

The strata, which are also the imputation classes in this example, are constructed according to company's size within each industry.

There are various variables in this survey. We consider the estimation of the population totals of the current year annual revenue (y) in four industries. The variable y has nonrespondents. Three covariates without nonrespondents are considered: the current year annual payroll, the previous year annual revenue, and the previous year annual payroll. The sample size, response size for y , and the sampling weight in each stratum and industry are given in Table 1.

Table 1
Sample Sizes, Response Sizes, and Sampling Weights
Across Industries and Strata

Industry	Stratum	Sample Size	Response Size	Sampling Weight
1	0	31	24	1.00
	1	14	6	12.43
	2	11	7	8.91
	3	10	4	6.10
	4	11	6	5.73
	5	16	12	2.70
2	6	18	13	2.17
	0	86	82	1.00
	1	8	2	32.91
	2	13	10	9.85
	3	11	9	10.82
	4	12	10	6.08
3	5	13	10	3.60
	0	38	30	1.00
	1	14	9	87.91
	2	11	8	67.39
	3	13	10	44.48
	4	14	13	25.28
	5	16	13	15.57
	6	18	12	9.80
	7	15	11	6.23
4	8	15	14	4.68
	9	40	33	2.13
	0	28	23	1.00
	1	7	5	32.14
	2	13	6	16.75
	3	10	7	12.90
	4	14	12	7.00
	5	13	9	6.18
	6	11	7	4.70
	7	17	12	3.31
	8	19	14	1.89
	9	22	16	1.82

First, we use the previous year annual revenue as the covariate x in simple cold deck imputation and ratio imputation. The current year annual payroll and the previous year annual payroll are used as \tilde{y} and \tilde{x} , respectively. For four industries and three imputation methods, Table 2 lists the estimated totals, the proposed estimated MSE's for the estimated totals, the naive estimated MSE's for the estimated totals (obtained by treating imputed values as

observed data), and the MSE ratios (the proposed estimated MSE over the naive estimated MSE). Note that the proposed estimated MSE is the sum of v_1 and v_2 for the ratio and cold deck-ratio methods or the sum of v_1 , v_2 , and the squared estimated bias for the simple cold deck method. Values of v_1 and v_2 are also included in the table.

Table 2

Estimated Totals and MSE's When x = the Previous Year Annual Revenue, \tilde{y} = the Current Year Payroll, and \tilde{x} = the Previous Year Annual Payroll

Industry	Estimate	Method		
		Cold Deck	Cold Deck-Ratio	Ratio
1	Total	5.31×10^9	5.19×10^9	5.42×10^9
	v_1	7.73×10^{14}	8.46×10^{14}	2.60×10^{15}
	v_2	1.39×10^{15}	2.50×10^{15}	1.81×10^{15}
	Proposed MSE	2.30×10^{15}	3.34×10^{15}	4.40×10^{15}
	Naive MSE	7.73×10^{14}	8.46×10^{14}	2.46×10^{15}
	MSE Ratio	2.97	3.95	1.79
2	Total	1.66×10^{10}	1.63×10^{10}	1.67×10^{10}
	v_1	4.00×10^{15}	4.19×10^{15}	5.57×10^{16}
	v_2	6.03×10^{15}	2.88×10^{16}	6.54×10^{15}
	Proposed MSE	1.02×10^{16}	3.30×10^{16}	6.23×10^{16}
	Naive MSE	4.00×10^{15}	4.19×10^{15}	5.58×10^{16}
	MSE Ratio	2.54	7.87	1.12
3	Total	3.54×10^{10}	3.53×10^{10}	3.59×10^{10}
	v_1	1.32×10^{16}	1.80×10^{16}	1.94×10^{17}
	v_2	5.44×10^{16}	8.62×10^{16}	6.77×10^{16}
	Proposed MSE	6.97×10^{16}	1.04×10^{17}	2.62×10^{17}
	Naive MSE	1.32×10^{16}	1.80×10^{16}	1.87×10^{17}
	MSE Ratio	5.27	5.80	1.40
4	Total	1.27×10^{10}	1.22×10^{10}	1.30×10^{10}
	v_1	2.11×10^{16}	2.14×10^{16}	5.13×10^{15}
	v_2	3.91×10^{15}	8.26×10^{15}	5.06×10^{15}
	Proposed MSE	2.59×10^{16}	2.97×10^{16}	1.02×10^{16}
	Naive MSE	2.11×10^{16}	2.14×10^{16}	5.06×10^{15}
	MSE Ratio	1.23	1.39	2.01

Next, to see the effect of using a wrong covariate in using the simple cold deck method, we repeat the previous computations using the current year annual payroll as the covariate x , and the previous year annual revenue and payroll as \tilde{y} and \tilde{x} , respectively. The results are reported in Table 3.

The following is a summary of the results in Tables 2 and 3.

1. The simple cold deck method depends heavily on the choice of the covariate x . When x is the previous year annual revenue (Table 2), the difference among the estimated totals provided by three methods is negligible; in terms of the estimated MSE, the simple cold deck method is the best. However, when x is the current year annual payroll (Table 3), the estimates from the simple cold deck is obviously too low; in terms of the estimated MSE, the simple cold deck method is the worst, because of its large bias (shown in Table 3).

Table 3

Estimated Totals and MSE's When x = the Current Year Annual Payroll, \tilde{y} = the Previous Year Annual Revenue, and \tilde{x} = the Previous Year Annual Payroll

Industry	Estimate	Method		
		Cold Deck	Cold Deck-Ratio	Ratio
1	Total	4.49×10^9	5.19×10^9	5.39×10^9
	Bias	-8.99×10^8		
	v_1	8.10×10^{14}	8.46×10^{14}	2.85×10^{15}
	v_2	1.38×10^{15}	2.64×10^{15}	1.75×10^{15}
	Proposed MSE	1.03×10^{16}	3.49×10^{15}	4.60×10^{15}
	Naive MSE	8.10×10^{14}	8.46×10^{14}	2.55×10^{15}
2	MSE Ratio	12.68	4.12	1.81
	Total	1.59×10^{10}	1.63×10^{10}	1.71×10^{10}
	Bias	-1.21×10^9		
	v_1	4.36×10^{15}	4.19×10^{15}	5.74×10^{16}
	v_2	8.20×10^{15}	1.48×10^{16}	8.95×10^{15}
	Proposed MSE	2.73×10^{16}	1.90×10^{16}	6.64×10^{16}
3	Naive MSE	4.36×10^{15}	4.19×10^{15}	5.62×10^{16}
	MSE Ratio	6.25	4.54	1.18
	Total	3.10×10^{10}	3.53×10^{10}	3.47×10^{10}
	Bias	-3.62×10^9		
	v_1	1.25×10^{16}	1.80×10^{16}	2.30×10^{17}
	v_2	4.56×10^{16}	9.25×10^{16}	5.41×10^{16}
4	Proposed MSE	1.89×10^{17}	1.10×10^{17}	2.84×10^{17}
	Naive MSE	1.25×10^{16}	1.80×10^{16}	1.83×10^{17}
	MSE Ratio	15.13	6.15	1.56
	Total	1.06×10^{10}	1.22×10^{10}	1.20×10^{10}
	Bias	-1.35×10^9		
	v_1	1.93×10^{16}	2.14×10^{16}	5.84×10^{15}
5	v_2	2.67×10^{15}	4.62×10^{15}	3.07×10^{15}
	Proposed MSE	4.03×10^{16}	2.60×10^{16}	8.92×10^{15}
	Naive MSE	1.93×10^{16}	2.14×10^{16}	8.92×10^{15}
	MSE Ratio	2.09	1.22	1.72

2. There is no definite conclusion on the relative performance (in terms of the estimated MSE) of the ratio imputation method and the cold deck-ratio method. In this example, the cold deck-ratio is better for industries 1-3, whereas the ratio imputation method is better for industry 4. Some scatter plots of the data (not shown) indicate that the correlation between x and \tilde{x} in industries 1-3 is higher than that in industry 4, which might be the reason for the difference in relative performance of the two imputation methods. See also the discussion after formula (12).
3. The naive estimated MSE's are much lower than the proposed estimated MSE's and are too optimistic. For example, in Table 3, the naive MSE's for the simple cold deck method are always smaller than those for the cold deck-ratio method, although we know that the simple cold deck does not work well in this case. In this example, v_2/v_1 is not small because of some large sampling fractions. Since the naive estimated MSE is either equal to v_1 (for the cold deck imputation

methods) or not very different from v_1 (for ratio imputation), the underestimation in using the naive estimated MSE is mainly due to treating imputed values as observed values in strata with large sampling fractions (and ignoring the bias of the simple cold deck estimators in the case of Table 3).

ACKNOWLEDGEMENT

The author would like to thank referees for helpful comments and suggestions. The first draft of this paper was finished at the U.S. Census Bureau and the U.S. Bureau of Labor Statistics when the author was an ASA/NSF Senior Research Fellow. The research was also supported by National Science Foundation Grants DMS-9504425 and DMS-9803112 and National Security Agency Grant MDA904-99-1-0032.

APPENDIX

1. **Proof of (7):** When $n/N \approx 0$, $V(\hat{Y}_R - Y) \approx V(\hat{Y}_R)$. Then (7) follows from

$$\begin{aligned} V(\hat{Y}_R) &= \frac{N^2}{n^2} \left\{ \sigma^2 E \left[\left(\sum_{i \in S} x_i \right)^2 / \left(\sum_{i \in r} x_i \right) \right] + \beta^2 V \left(\sum_{i \in S} x_i \right) \right\} \\ &\approx \frac{N^2}{n} \left(\frac{\sigma^2 \mu_x}{p} + \beta^2 v_x \right) \end{aligned}$$

for large n , where the last approximate equality follows from the fact that conditioned on x_i 's, $E(\sum_{i \in r} x_i) = p \sum_{i \in S} x_i$.

2. **Proof of (9):** Under model (1),

$$\begin{aligned} V(\hat{Y}_C) &= \frac{N^2}{n^2} \left\{ V \left(\sum_{i \in r} x_i^{1/2} e_i \right) + V \left(\beta \sum_{i \in r} x_i + \sum_{i \in S-r} x_i \right) \right\} \\ &= \frac{N^2}{n^2} \left\{ \sigma^2 p \mu_x + \beta^2 V \left(\sum_{i \in r} x_i \right) + V \left(\sum_{i \in S-r} x_i \right) \right. \\ &\quad \left. + 2\beta \text{Cov} \left(\sum_{i \in r} x_i, \sum_{i \in S-r} x_i \right) \right\} \\ &= \frac{N^2}{n} \{ \sigma^2 p \mu_x + \beta^2 [p v_x + p(1-p) \mu_x^2] \\ &\quad + (1-p)(v_x + p \mu_x^2) - 2\beta p(1-p) \mu_x^2 \} \\ &= \frac{N^2}{n} \{ \sigma^2 p \mu_x + (\beta^2 p + 1 - p) v_x \\ &\quad + (\beta - 1)^2 p(1-p) \mu_x^2 \}. \end{aligned}$$

3. **Proof of (11):** Under the assumed conditions on (y_i, x_i) and $(\tilde{y}_i, \tilde{x}_i)$,

$$\begin{aligned} \text{mse}(\hat{Y}_{C-R}) &= \frac{N^2}{n^2} V \left(\sum_{i \in r} y_i + \sum_{i \in S-r} z_i \right) \\ &= \frac{N^2}{n^2} V \left(\sum_{i \in r} \epsilon_i + \sum_{i \in S} z_i \right) \\ &= \frac{N^2}{n^2} \left\{ V \left(\sum_{i \in r} \epsilon_i \right) + V \left(\sum_{i \in S} z_i \right) \right. \\ &\quad \left. + 2\text{Cov} \left(\sum_{i \in r} \epsilon_i, \sum_{i \in S} z_i \right) \right\} \\ &= \frac{N^2}{n} \{ \sigma^2 p (\mu_x + \gamma_x) + (\beta^2 v_x + \sigma^2 \gamma_x) - 2\sigma^2 p \gamma_x \} \\ &= \frac{N^2}{n} \{ \sigma^2 p \mu_x + \beta^2 v_x + \sigma^2 (1-p) \gamma_x \}. \end{aligned}$$

REFERENCES

- BUTANI, S., HARTER, R., and WOLTER, K. (1998). Estimation procedures for the Bureau of Labor Statistics current employment statistics program. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.
- COCHRAN, W.G. (1977). *Sampling Techniques*. Third Edition. New York: Wiley.
- KALTON, G., and KASPRZYK, D. (1986). The treatment of missing data. *Survey Methodology*, 12, 1-16.
- KING, C., and KORNBAU, M. (1994). Inventory of economic area statistical practices. ESMD Report Series 9401, Bureau of the Census, Washington D.C.
- KREWSKI, D., and RAO, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.
- LEE, H., RANCOURT, E., and SÄRNDAL, C.-E. (1994). Experiments with variance estimation from survey data with imputed values. *Journal of Official Statistics*, 10, 231-243.
- RAO, J.N.K., and SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.
- RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- VALLIANT, R. (1993). Poststratification and conditional variance estimation. *Journal of American Statistical Association*, 88, 89-96.

Model-Based Estimation With Link-Tracing Sampling Designs

STEVEN K. THOMPSON and OVE FRANK¹

ABSTRACT

Samples from hidden and hard-to-access human populations are often obtained by procedures in which social links are followed from one respondent to another. Inference from the sample to the larger population of interest can be affected by the link-tracing design and the type of data it produces. The population with its social network structure can be modeled as a stochastic graph with a joint distribution of node values representing characteristics of individuals and arc indicators representing social relationships between individuals. In this paper maximum likelihood estimators of population graph parameters are described. Predictors of realized population graph quantities are obtained using predictive likelihood. These estimators and predictors are compared with conventional data summaries and illustrated with a numerical example.

KEY WORDS: Snowball samples; Adaptive sampling; Graph sampling; Ignorable designs; Link-tracing designs; Network sampling; Likelihood; Predictive likelihood.

1. INTRODUCTION

In studies of hidden and hard-to-access human populations, link-tracing procedures, in which social links are followed from one respondent to another, are commonly involved in obtaining the sample. For example, in a study of injection drug use in relation to the spread of the HIV infection, initial respondents may be asked to identify drug-injection or sexual partners who are then added to the sample. For such a study, the social links are of inherent importance for understanding the issues of interest while at the same time being useful or essential in building the sample. However, inference from the sample to the larger population or social structure of interest can be affected by the link-tracing procedures and the type of data they produce. In this paper we evaluate this inference problem in relation to the design and describe some inference methods for such studies based on maximum likelihood estimation and prediction.

Human populations with social structure are often modeled as graphs, with the nodes of the graph representing individuals and the edges or arcs of the graph representing social links, relationships, or transactions. The population graph itself can be viewed either as a fixed structure or as a realization of a stochastic graph model. In real studies of human populations, particularly those that are hidden or hard to access, it is seldom possible to obtain data on the whole population or graph structure. Rather, data are obtained from a sample, and the sample may have been obtained by innovative and unconventional means, including methods taking advantage of the arcs or links from one individual to another. The data may contain information about characteristics of sample individuals, social links within the sample, and in some cases information about links between individuals in the sample and individuals outside the sample.

In this paper we use the term "sampling design" to refer to the procedure by which the sample is selected, whether deliberate or happenstance. For many ethnographic and sociological studies of hidden populations, link-tracing designs are considered the only practical way to obtain a sample large enough to study. In other studies, the social structure is itself the object of interest and the link-tracing methods are used in order to obtain meaningfully structured samples to study.

The statistical literature on design and estimation with link-tracing designs includes procedures variously termed snowball sampling, chain-referral sampling, random walks, nexus sampling, network or multiplicity sampling, and adaptive sampling. A type of link-tracing design in which individuals in an initial sample were asked to identify a fixed number of acquaintances, who in turn were asked to identify the same number of acquaintances, and so on for a fixed number of stages or waves, was termed "snowball sampling" by Goodman (1961). A Bernoulli procedure was assumed for the initial sample. Snowball designs were developed in the graph setting with a variety of initial probability sampling designs and any numbers of links and waves by Frank (1971, 1977a,b, 1978a,b, 1979a), who obtained a variety of design and model based methods for estimating graph quantities from the sample data. Snijders (1992) used the same term "snowball sampling" to include designs in which only a subsample of links from each node is traced. The case in which only one of the links from a node is selected at random and followed to another node, and then one of its links selected, and so on, was called a "random walk" by Klovdahl 1989. Link-tracing sampling methods in which there is only one link from each node have been termed "chains" (Erickson 1979). Frank and Snijders (1994) consider model- and design-based

¹ Steven K. Thompson, Department of Statistics, 326 Thomas Building, Pennsylvania State University, University Park, PA 16802 USA; Ove Frank, Department of Statistics, Stockholm University, S-10691 Stockholm, Sweden. This research is part of an ongoing, equal collaboration effort and order of authorship was determined by a coin toss.

estimation of a hidden population size, that is, the number of nodes in the graph, based on snowball samples. Additional practical and statistical issues in sampling from social networks with various types of snowball, chain-referral, and other link-tracing designs are discussed in Granovetter (1976), Morgan and Rytina (1977), Frank (1979b, 1981, 1988), Watters and Biernacki (1989), van Meter (1990), Spreen (1992), Wasserman and Faust (1994), Spreen and Zwaagstra (1994), Karlberg (1997), Jansson (1997), Spreen (1998), and Robins (1998).

Design-based estimation methods were developed additionally for the closely related designs of network or multiplicity sampling, in which social, kinship, and administrative links were traced (Birnbaum and Sirken 1965, Kalton and Anderson 1986, Levy 1977, Levy and Lemeshow 1991, Sirken 1970, 1972a, b, Sirken and Levy 1974, Sudman, Sirken, and Cowan 1988). For example, in a survey of a rare disease, an initial sample of households might be selected at random and data obtained both for residents of the households and for their siblings. The design-based estimation in these strategies is helped by the symmetry of the links and the encompassing of complete connected components in the sample, and unbiased estimators have been obtained for network sampling with many different initial designs.

Another link-tracing procedure for which design-based estimators are available is adaptive cluster sampling (Thompson 1990, 1997, Thompson and Seber 1996), which has been formulated in the graph setting as well as the spatial setting. Following selection of an initial sample of nodes by any of a number of initial designs, the decision on whether to follow links from a node or not depends on the value of a variable of interest observed for the node. For example, in an epidemiological study of a sexually transmitted disease, sexual or social links may be followed only from respondents who have been infected. Design-unbiased estimation methods have been worked out for a wide variety of adaptive cluster sampling strategies.

Design-based methods of inference, such as the design-based estimation methods of network sampling, snowball sampling, and adaptive cluster sampling, have the advantage that properties such as design-unbiasedness or consistency do not depend for their validity on any assumed model for the population. On the other hand, these properties do depend on the sampling design being carried out as specified. The model-based methods described in this paper, on the other hand, do depend on an assumed model for the population or graph. Their practical advantage is that they apply to a wide range of sample selection procedures, and thus allow more leeway in how the sample is actually selected.

In fact many real studies of hidden and hard-to-reach populations use sample selection procedures, including link-tracing, that are not readily analyzed based on idealized design-induced probabilities. For example, in a study to examine the relation of network structure and risk behaviors

such as needle sharing among drug injectors in the Bushwick section of Brooklyn, "index" (initial) respondents were used as "auxiliary recruiters" to bring members of their networks into the study (Friedman, Neaigus, Jose, Curtis, Goldstein, Ildefonso, Rothenberg and Des Jarlais 1997, Neaigus, Friedman, Goldstein, Ildefonso, Curtis and Jose 1995, Neaigus, Friedman, Jose, Goldstein, Curtis, Ildefonso and Des Jarlais 1996). Only about 61% of the linked individuals were actually recruited, however. In a long-term study on the heterosexual transmission of HIV infection (Rothenberg, Woodhouse, Potterat, Muth, Darrow and Klovdahl 1995), the target population of interest consisted of commercial sex workers, their paying and nonpaying partners, persons who use injectable drugs, and the sexual partners of drug users in the Colorado Springs area. Persons in the purposively-selected initial sample were interviewed and, in addition to their individual characteristics, identities of their sexual partners were obtained. Persons named by two or more respondents were also located and interviewed. The wide range of link-tracing procedures used in studies such as these has motivated the emphasis in this paper on model-based inference methods.

When we compare the maximum likelihood estimators and predictors obtained in this paper with commonly-used conventional estimates or data summaries such as sample means and proportions of node or link values, we find that in most cases the conventional estimates are not the best estimates. Similarly, estimators that would be appropriate if the data included the whole graph may not be appropriate with data on only a sample from the graph. An implication of these results is that conventional estimates or unadjusted summaries of sample data obtained through link-tracing procedures can be misleading if viewed as pertaining to population or whole-graph characteristics. The interpretations of this discrepancy provided in this paper give some insight into the conditions under which the best estimate would tend to be lower, or higher, than the conventional one.

Notation and basic issues for design and inference in the graph setting are presented in section 2. In section 3, a wide range of link-tracing procedures, all of which can be analyzed using the approach presented in this paper, are described. In section 4, a class of graph models that we use to illustrate the inference methods of the paper is described. Estimative and predictive maximum likelihood methods for graph parameters and realized population values are described in section 5.

2. GRAPH MODELS AND SAMPLING DESIGNS

Consider a graph of N nodes (units) labeled $1, 2, \dots, N$. Associated with the u -th node is a variable of interest Y_u . We denote the full set of node labels $U = \{1, 2, \dots, N\}$ and the sequence of node variables by $\mathbf{Y} = (Y_1, \dots, Y_N)$. For two distinct nodes u and v , the indicator variable X_{uv} equals

one if there is an arc (directional link) from u to v and zero otherwise. The matrix of arc indicators, having X_{uv} as the element in the u -th row and v -th column, is the graph adjacency matrix, denoted \mathbf{X} . For convenience we will assume the diagonal elements X_{uu} are zero. The ordered pair (u, v) is sometimes referred to as a dyad of type $(Y_u, Y_v; X_{uv}, X_{vu})$. A graph model is given by a joint probability or density $f(\mathbf{y}, \mathbf{x}; \psi)$ for outcomes \mathbf{y} and \mathbf{x} of \mathbf{Y} and \mathbf{X} , respectively, and it may depend on one or more unknown parameters ψ .

A sample s from the graph is a subset of nodes and a subset of node pairs. We can write the combined sample as $s = (s^{(1)}, s^{(2)})$, where $s^{(1)}$ denotes the subset of nodes selected for observation of the associated y -values and $s^{(2)}$ denotes the subset of node pairs selected for observation of the associated x -values. The data consist of the node and node-pair labels in the combined sample together with the associated node and arc-indicator values, that is $d = (u, (v, w), y_u, x_{vw}; u \in s^{(1)}, (v, w) \in s^{(2)})$ or, more simply, $d = (s, \mathbf{y}_s, \mathbf{x}_s)$. Further, it is often convenient to use \mathbf{y}_s to denote the y -values of the nodes in the combined sample and \mathbf{x}_s for the x -values of the node pairs in the combined sample, with $\mathbf{y}_{\bar{s}}$ and $\mathbf{x}_{\bar{s}}$ denoting the values of the unsampled nodes and node pairs. Often the sampling procedure results in a connection between $s^{(1)}$ and $s^{(2)}$. For example, if all relationships from sample nodes to other sample nodes, and no others, are recorded, then $s^{(2)} = s^{(1)} \times s^{(1)}$. In general, however, the nodes on which y -values are recorded and the node pairs on which x -values are recorded may be quite unrelated sets. In particular, the link-tracing procedures considered in this paper often lead to data on links from nodes in $s^{(1)}$ to nodes outside of $s^{(1)}$.

The sampling design is the procedure by which the sample is selected. This selection procedure may be controlled by the investigators, as is the case with a deliberately implemented probability sampling design, or may be beyond the control of the investigators and determined by the circumstances of the situation. If the probability of selecting the sample does not depend on node values y or link values x or parameters ψ involved in the graph model, we refer to the design as "conventional." For a conventional design the probability of selecting sample s can be written $p(s)$ or $p(s; \phi)$, where ϕ denotes any unknown parameters involved in the design (but not the model), as in a Bernoulli sampling with unknown inclusion probability ϕ for each node. The sampling design may depend on one or more auxiliary variables that are known for the whole population, but that dependence will be left implicit in the notation $p(s)$. Conventional designs include the classical probability designs such as simple random, systematic, stratified, multi-stage, and unequal probability sampling, as well as model-based purposive and balanced designs based on auxiliary variables.

If the probability of selecting the sample depends on any y or x values, we refer to the design as "adaptive," since the selection procedure adapts to the realized configuration of

node and link values in the population. In addition, the design can involve unknown parameters ψ . Thus, in general the sampling design in the graph setting has a selection probability that can be written $p(s | \mathbf{y}, \mathbf{x}; \psi)$ where \mathbf{y} denotes the sequence of node values, \mathbf{x} the matrix of arc indicator values, and ψ any parameters involved.

Likelihood-based inference, such as maximum likelihood estimation or prediction and Bayes methods, is simplified if the design can be ignored at the inference stage. The fact that the sampling design does not affect the value of a Bayes or likelihood-based estimator in survey sampling was noted by Godambe (1966) for designs that do not depend on any values of the variable of interest and by Basu (1969) for designs that do not depend on values of the variable of interest outside the sample. Scott and Smith (1973) showed that the design could become relevant to inference when the data lacked information about the labels of the units in the sample. Rubin (1976) gave exact conditions for a missing data mechanism – of which a sampling design can be viewed as an example – to be relevant in frequentist and likelihood-based inference. For likelihood-based methods such as maximum likelihood and Bayes methods, the design is "ignorable" if the design or mechanism does not depend on values of the variable of interest outside the sample or on any parameters in the distribution of those values. For frequency-based inference such as design- or model-unbiased estimation, however, the design is relevant if it depends on any values of the variable of interest, even in the sample. Scott (1977) showed that the design is relevant to Bayes estimation if auxiliary information used in the design is not available at the inference stage. Sugden and Smith (1984) gave general and detailed results on when the design is relevant in survey sampling situations. Thompson and Seber (1996) described adaptive designs in which the selection procedure deliberately takes advantage of observed values of the variable of interest, and discussed the relevance of the design in inference from a variety of design and model based perspectives. Similar issues of design and inference arise with adaptive experimental designs, such as medical experiments in which ethical considerations motivate adaptive treatment allocation to favor the more promising treatments as the study progresses (*cf.* Flournoy and Rosenberger 1995, Rosenberger 1996, Wei, Smythe, Lin and Park 1990). It is important to underscore that a design that is said to be "ignorable" for likelihood-based inference might not be ignorable for a frequentist-based inference, such as model-unbiased estimation, and that even though a design may be ignorable at the inference stage, in that for example the way an estimator is calculated does not depend on the design used, the design is still relevant *a priori* to the properties of the estimator.

The sample data $d = (s, \mathbf{y}_s, \mathbf{x}_s)$ are a function of the sample selected and of the graph values \mathbf{y} and \mathbf{x} . The likelihood can be written

$$L(\psi, d) = \sum p(s | \mathbf{y}, \mathbf{x}; \psi) f(\mathbf{y}, \mathbf{x}; \psi) \quad (1)$$

where the sum is over outcomes (\mathbf{y}, \mathbf{x}) consistent with the data d . Since the y and x values for nodes and node pairs in the sample are fixed by the data, the sum is over all possible values of the unobserved variables $\mathbf{y}_{\bar{s}}$ and $\mathbf{x}_{\bar{s}}$ and it actually represents the marginal probability of the sample s selected and the associated observed variables \mathbf{y}_s and \mathbf{x}_s .

Thus, in general the likelihood function depends on both the design and the model. The quantity $\sum_{\mathbf{y}_{\bar{s}}, \mathbf{x}_{\bar{s}}} f(\mathbf{y}, \mathbf{x}; \psi)$, based on the model only without consideration of the design, was termed the “face-value likelihood” by Dawid and Dickey (1977) because inference based on this function alone takes the data at face value without considering how the data were selected.

For any design in which the selection of the sample depends on graph y and x values only through those values \mathbf{y}_s and \mathbf{x}_s included in the data, the design probability can be moved out of the sum and forms a separate factor in the likelihood. If in addition the design and model parameters are distinct and not related, the likelihood can be written

$$L(\phi, \psi, d) = p(s | \mathbf{y}_s, \mathbf{x}_s; \phi) \sum_{\mathbf{y}_{\bar{s}}, \mathbf{x}_{\bar{s}}} f(\mathbf{y}, \mathbf{x}; \psi) \quad (2)$$

where ϕ denotes the design parameters and ψ denotes the model parameters. The design then does not affect the value of estimators or predictors based on direct likelihood methods such as maximum likelihood or Bayes estimators. For any such “ignorable” design, the sum in the above likelihood, over all values of \mathbf{y} and \mathbf{x} leading to the given data value, is simply the marginal probability of the y and x values associated with the sample data. This marginal distribution depends on what sample was selected, but does not depend on how that sample was selected. For likelihood-based inference with a design ignorable in this sense, the face-value likelihood gives the correct inference.

3. SOME LINK-TRACING DESIGNS

A variety of link-tracing designs are described in this section. Each of these designs is ignorable in the likelihood sense provided the initial sample is selected by an ignorable procedure and provided the data include all the values involved in the selection procedure. Since for all the designs described in this section, the node-pair sample $s^{(2)}$ has a deterministic functional relationship to the node sample $s^{(1)}$, the superscript notation will be omitted and the final node sample $s^{(1)}$ will be denoted simply s .

The simple likelihood methods described in this paper apply to a wide range of ignorable link-tracing designs, including those described in this section. Further research is needed on methods for nonignorable designs, including those with nonignorable selection of the initial sample. Methods for dealing with nonsampling errors such as non-response and reporting errors with link-tracing designs are also in need of further development (cf., Thompson 1997).

3.1 Single-Wave Design

In a single-wave link-tracing design an initial sample of nodes is selected by any ignorable design from the population of nodes in the graph. For each node in the sample, nodes adjacent from that node are added to the sample. The snowball procedure is assumed to stop after one wave. Thus, node v will be added if for some node u in the initial sample $x_{uv} = 1$.

Let s_0 denote the set of nodes in the initial sample and s_1 denote the added nodes not in the initial sample. The whole sample is $s = s_0 \cup s_1$.

The entire set of labels can be written as the union of three disjoint sets, $U = s_0 \cup s_1 \cup \bar{s}$. The values y associated with the nodes can be correspondingly ordered as a sequence $(\mathbf{y}_{s_0}, \mathbf{y}_{s_1}, \mathbf{y}_{\bar{s}})$, where $\mathbf{y}_a = (y_u; u \in a)$ is the subsequence of y restricted to indices in subset $a \subset U$. The adjacency matrix \mathbf{x} is ordered correspondingly and partitioned into submatrices $\mathbf{x}_{s_0 s_0}, \mathbf{x}_{s_0 s_1}, \mathbf{x}_{s_0 \bar{s}}$ and so on, where $\mathbf{x}_{ab} = (x_{uv}; u \in a, v \in b)$. Ordering the adjacency matrix in this way facilitates the specification of factors in the likelihood.

With the design above, the probability of selecting sample s depends only on $\mathbf{x}_{s_0 U}$ and so can be written $p(s | \mathbf{x}_{s_0 U})$, where $\mathbf{x}_{s_0 U}$ can also be replaced by its column permutation $(\mathbf{x}_{s_0 s_0}, \mathbf{x}_{s_0 s_1}, \mathbf{x}_{s_0 \bar{s}})$. That is, the probability of selecting the final sample $s = s_0 \cup s_1$ depends on links from the initial sample to other units in the graph, both in s and in \bar{s} . The data consist of $(s, \mathbf{y}_s, \mathbf{x}_{s_0 U})$. Since the design does not depend on any x or y values outside the data or on model parameter values, the design is ignorable for likelihood-based inference.

3.2 Multi-Wave Samples

Consider a snowball sample with $k + 1$ waves after the initial sample. The sample will be denoted $s = s_0 \cup s_1$ with $s_0 = s_{00} \cup s_{01} \cup s_{02} \cup \dots \cup s_{0k}$. An initial sample s_{00} is selected by any design that is ignorable in the likelihood sense. Links are followed and every node with an arc from any node in s_{00} and not already in the sample is added to form the first-wave sample s_{01} . That is, $s_{01} = \{v: x_{uv} = 1 \text{ for some } u \in s_{00}, v \notin s_{00}\}$. Then links are followed in s_{01} to give the second-wave sample $s_{02} = \{v: x_{uv} = 1 \text{ for some } u \in s_{01}, v \notin s_{00} \cup s_{01}\} = \{v: x_{uv} = 1 \text{ for some } u \in s_{00} \cup s_{01}, v \notin s_{00} \cup s_{01}\}$. Finally, the $(k + 1)$ -wave sample, denoted simply s_1 , is added by following links from the k -th wave sample s_{0k} . That is $s_1 = \{v: x_{uv} = 1 \text{ for some } u \in s_{0k}, v \notin s_{0k}\}$. No links from s_1 are followed.

If $s_{0j} = \emptyset$ for any $j < k$ then sampling stops, so that the number of waves added is less than k if at some point there are no links leading out of the current sample to unsampled nodes.

The data consist of sets of node labels in the different waves of the sample and the ordered node pairs from s_0 to U , the sequence of node-values y_s for all nodes in the sample, and the link indicator variables $\mathbf{x}_{s_0 U}$ from s_0 to the set U of nodes in the graph. Thus the data consist of the

subgraph data for s_0 , that is $(s_0, \mathbf{y}_{s_0}, \mathbf{x}_{s_0s_0})$, together with the node values \mathbf{y}_{s_1} for the nodes in the final-wave s_1 , the link indicators $\mathbf{x}_{s_0s_1}$ from s_0 to s_1 , and the link indicators $\mathbf{x}_{s_0\bar{s}}$ from the nodes in s_0 to the nodes not in the sample.

Since the design does not depend on any y or x values outside the data nor on any of the graph model parameters, the design is ignorable and the structure of the data is exactly the same with the $(k+1)$ -wave snowball as with the 1-wave snowball design, and with the notation we have used the likelihood and estimation formulas are unchanged with the more general design.

3.3 Completed-Wave Designs

With a completed snowball sample, the procedure of adding waves is continued until no further links lead out of the sample. Then the number of completed waves K is a random variable and $s_{0,K+1} = s_1$ is the first empty set in the sequence (s_{00}, s_{01}, \dots) . The data are $d = (s_0, \mathbf{y}_{s_0}, \mathbf{x}_{s_0U})$ or equivalently $(s_0, \mathbf{y}_{s_0}, \mathbf{x}_{s_0s_0}, \mathbf{x}_{s_0\bar{s}_0})$. Inference can then proceed with the same likelihood and estimation formulas but with the simplification that the data contains no set s_1 for which \mathbf{y}_{s_1} and \mathbf{x}_{s_0,s_1} are known but from which links are unknown.

3.4 Link-Tracing Adaptive on Node Values

Consider a design in which the decision to follow the links from node u depends on the node value y_u . For example, in a study on injection drug use, the initial sample may contain both users ($y_u = 1$) and nonusers ($y_u = 0$). If the investigators choose to follow social links only from users, then the design depends adaptively on the node y -values as well as the links. Similarly, in a study of sexually transmitted diseases, investigators may be instructed to follow sexual or social links more frequently from infected respondents than from noninfected respondents. The design then can be written $p(s|\mathbf{y}_s, \mathbf{x}_{s_0U})$, since the selection procedure depends on both node and link values. If the data contain all values on which the design depends, that is, $d = (s, \mathbf{y}_s, \mathbf{x}_{s_0U})$, then the design is ignorable and maximum likelihood inference is simplified as described in the following sections.

3.5 Tracing Only a Subsample of Sample Links

The designs described above can be generalized to procedures in which only a sample of the links leading out from node u in s_0 are followed. Examples include the "random walk" design of Klov Dahl (1989) and the generalization of snowball designs described in Snijders (1992). In the random walk design, an initial respondent is asked to give the names of several social contacts. One of these contacts is chosen at random to be interviewed and asked in turn to name several contacts, one of which is chosen at random, and so on. In practice, dead ends can occur when a respondent either reports no contacts or reports only contacts who are already in the sample. In such cases investigators either backtrack and try different

leads from previous respondents or find a new initial respondent.

With these subsampling link-tracing designs, the procedure for selecting the sample, though complicated from a design-probability point of view, depends only on values in the sample and on links leading from the sample. We again assume that the initial sample is obtained by any ignorable procedure. Let $s_0 = s_{00} \cup s_{01} \cup s_{02} \cup \dots \cup s_{0k}$ consist of all of the waves from which at least some links are followed. Thus, s_{01} consists of the nodes not previously included obtained by following a subsample of the links from nodes in the initial sample s_{00} , s_{02} consists of the nodes not previously included obtained by following a subsample of the links from nodes in $s_{00} \cup s_{01}$, and so on. No links are followed from the final wave s_1 . Allowing for the possibility of dependence on node values, the design can be written $p(s|\mathbf{y}_s, \mathbf{x}_{s_0U})$, so that with data $d = (s, \mathbf{y}_s, \mathbf{x}_{s_0U})$, the design is ignorable for likelihood-based inference.

3.6 Data from Link-Tracing Designs

With any of the single or multi-wave link-tracing designs described above, it is of considerable practical importance what data are recorded. If the data include only the sample node labels, the y -values for nodes in the sample, and the arc indicators for pairs of units in the sample, that is, $d = (s, \mathbf{y}_s, \mathbf{x}_{ss})$, then the design is nonignorable and must be integrated into the likelihood, which can complicate analysis.

Consider also a study in which social links are used in the design, to find the sample, but only node characteristics (y -values), not relationships are recorded, so that the data are $d = (s, \mathbf{y}_s)$. Then the design is nonignorable.

If on the other hand the data from the link-tracing design include not only the linkages within the sample but the out-linkages (or lack thereof) from all but the last wave to the rest of the graph, that is, $d = (s, \mathbf{y}_s, \mathbf{x}_{s_0U})$, then the design depends only on graph values in the data and so factors out of the likelihood.

4. A GRAPH MODEL WITH LINKS RELATED TO NODE VALUES

The likelihood-based approach described in section 2 with sample data from link-tracing designs of types described in section 3 will be illustrated using a class of graph models described in this section. This class of models builds on conditional independence between dyads as in the contact models of Frank (1979a) and Wellman, Frank, Espinoza, Lundquist and Wilson (1991). Conditional on the node values, independence is assumed between dyads, with the distribution of links between pairs of nodes depending on node value. Thus, unconditionally these models have dependence between dyads because of the dependence on the node values. In the models of Holland

and Leinhardt (1981), dyads are assumed to be independent but with distributions that depend on fixed node parameters. Wasserman (1980) also assumed independence of dyads in modeling the change in a graph over time as a stochastic process. Bayesian extensions and stochastic blockmodels of Holland, Laskey, and Leinhardt (1983), Fienberg, Meyer, and Wasserman (1985), Wang and Wong (1987), and Frank (1988) provide generalizations to joint distributions with dependence between node values and graph links. Models by Frank and Harary (1982) for randomly colored graphs exhibit a similar structure. The Markov graph models of Frank and Strauss (1986) provide another approach to dependence among dyads but present difficulties for maximum likelihood estimation. Review of a variety of graph models is found in Wasserman and Faust (1994) and Frank (1997).

The maximum likelihood estimation and prediction methods of this paper apply equally to sample data with graph models other than the class of stochastic block models we have used. With other models, the same conditions for ignorability apply. We have chosen this class of models because it is rich enough to encompass important aspects of realism such as dependence between dyads and between dyads and node values, and it is simple enough to have explicit full-graph maximum likelihood estimators for comparison with the estimators based on samples. With other classes of models such as the Markov graph models, estimation even with full-graph data requires numerical methods.

For practical use of the model based approach it is important to have diagnostic tools for evaluations and comparisons between alternative models. For example, with the two-block model used here the conditional independence of dyads could be tested by counting pairs of dyads of different types within and between the blocks. Within each block there are three types of dyads and six types of pairs of dyads. Between the two blocks there are four types of dyads and ten types of pairs of dyads. A Pearson goodness-of-fit statistic between observed and expected counts of the 22 types of pairs of dyads within and between the blocks is asymptotically chi-square distributed with 12 degrees of freedom under the conditional dyad independence assumption. Goodness-of-fit testing for graph models is discussed by Holland and Leinhardt (1981) and Frank and Strauss (1986), and this direction of research needs further development in particular in connection with sample data from link-tracing designs.

In the assumed model the node variables Y_1, \dots, Y_N are independent, identically distributed (i.i.d.) Bernoulli random variables with probabilities $P(Y_u = i) = \theta_i$, for $i = 0, 1$, with $\theta_0 + \theta_1 = 1$. Conditional on the node values Y_1, \dots, Y_N , the dyads (X_{uv}, X_{vu}) are independent, for $1 \leq u < v \leq N$, with conditional distribution given by $P[(X_{uv}, X_{vu}) = (k, l) | Y_u = i, Y_v = j] = \lambda_{ijkl}$ for all combinations of $i = 0, 1, j = 0, 1, k = 0, 1$, and $l = 0, 1$. For all combinations of i and j , the sums over k and l are denoted

$\lambda_{ij..} = \sum_k \sum_l \lambda_{ijkl}$ and equal 1. In order to get graph probabilities not depending on node identities, the following symmetry requirements are needed: $\lambda_{1110} = \lambda_{1101}$, $\lambda_{1011} = \lambda_{0111}$, $\lambda_{1010} = \lambda_{0101}$, $\lambda_{1001} = \lambda_{0110}$, $\lambda_{0010} = \lambda_{0001}$, and $\lambda_{1000} = \lambda_{0100}$. The pattern of these restrictions is illustrated in Table 1.

Table 1

		(x_{uv}, x_{vu})			
(y_u, y_v)		(0,0)	(0,1)	(1,0)	(1,1)
(0,0)		●	● — ●	●	●
(0,1)		●	● — ●	●	●
(1,0)		●	● — ●	●	●
(1,1)		●	● — ●	●	●

With these restrictions, it is convenient to introduce the notation

$$\lambda_{ijkl} = \begin{cases} \gamma'_{i+j, k+l}, & \text{if } (ijkl) = (0110) \text{ or } (1001), \\ \gamma_{i+j, k+l}, & \text{otherwise} \end{cases}$$

where $\gamma_{00} + 2\gamma_{01} + \gamma_{02} = 1$, $\gamma_{10} + \gamma_{11} + \gamma'_{11} + \gamma_{12} = 1$, and $\gamma_{20} + 2\gamma_{21} + \gamma_{22} = 1$. We can interpret γ'_{11} and γ_{11} as the probabilities of dyads with an arc from an unmarked to a marked node only and from a marked to an unmarked node only, respectively. Moreover, for $(ij) \neq (11)$, γ_{ij} is the probability of a dyad with j arcs on i marked and $2 - i$ unmarked nodes.

It will also be convenient to denote $\lambda_{ij1.} = \sum_l \lambda_{ij1l} = \alpha_{ij}$ and $\lambda_{ij11} = \beta_{ij}$ for $i = 0, 1$ and $j = 0, 1$. Here α_{ij} is the probability of an arc from a node of value i to a node of value j , and β_k is the probability of mutual arcs between k marked nodes.

Let N_i denote the total number of nodes with value i in the graph, for $i = 0, 1$, so that $N_0 + N_1 = N$. Let further M_{ijkl} denote the total number of dyads of type $(ijkl)$, that is, the total number of ordered node pairs (u, v) , with $u < v$, such that $(Y_u, Y_v, X_{uv}, X_{vu}) = (ijkl)$.

The likelihood for the full graph under the model with parameters (θ, λ) is

$$L(\theta, \lambda; \mathbf{y}, \mathbf{x}) = \left(\prod_{i=0}^1 \theta_i^{N_i} \right) \left(\prod_{i=0}^1 \prod_{j=0}^1 \prod_{k=0}^1 \prod_{l=0}^1 \lambda_{ijkl}^{M_{ijkl}} \right). \quad (3)$$

In terms of the γ s,

$$\prod_{i=0}^1 \prod_{j=0}^1 \prod_{k=0}^1 \prod_{l=0}^1 \lambda_{ijkl}^{M_{ijkl}} = \left(\prod_{i=0}^1 \prod_{j=0}^1 \gamma_{ij}^{R_{ij}} \right) (\gamma'_{11})^{R'_{11}}$$

where the R s are dyad counts corresponding to the pattern in Table 1. That is, $R_{00} = M_{0000}$, $R_{01} = M_{0001} + M_{0010}$,

$R_{02} = M_{0011}$, $R_{10} = M_{0100} + M_{1000}$, $R_{11} = M_{0101} + M_{1010}$, $R'_{11} = M_{0110} + M_{1001}$, $R_{12} = M_{0111} + M_{1011}$, $R_{20} = M_{1100}$, $R_{21} = M_{1101} + M_{1110}$, $R_{22} = M_{1111}$. Note that R'_{11} (R_{11}) is the number of dyads with an arc from an unmarked (marked) to a marked (unmarked) node only. Also note that except for $(ij) = (11)$, R_{ij} is the number of dyads on i marked nodes with j arcs.

The maximum likelihood estimators with the whole graph as data are the proportions $\hat{\theta}_i = N_i / N$, $\hat{\gamma}_{ij} = R_{ij} / R_i$, and $\hat{\gamma}'_{11} = R'_{11} / R_1$, where $R_0 = N_0(N_0 - 1) / 2$, $R_1 = N_0N_1$, and $R_2 = N_1(N_1 - 1) / 2$. In terms of the λ s, this means $\hat{\lambda}_{ijkl} = R'_{11} / R_1$ if $(ijkl) = (0110)$ or (1001) and $\hat{\lambda}_{ijkl} = R_{i+j,k+l} / R_{i+j}$ otherwise.

5. INFERENCE FROM LINK-TRACING DESIGNS

5.1 Estimating Graph Model Parameters

Consider any of the link-tracing designs, for which an initial or multiwave sample is selected and links out from nodes in s_0 are followed to add the set s_1 of nodes not in s_0 that are adjacent after nodes in s_0 . The data are $d = (s, y_s, x_{s_0U})$, so that the design depends on y and x values only through those in the data and is thus ignorable.

With the graph model described in the previous section, the likelihood with the sample data given by equation (2) in section 2 is in this case

$$L(\theta, \lambda, d) = P(s | y_s, x_{s_0U}) \sum \left(\prod_{u=1}^N \theta_{y_u} \right) \left(\prod_{u < v} \lambda_{y_u y_v x_{uv} x_{vu}} \right)$$

where the sum is over all values y_u and x_{uv} that are not fixed by the sample data.

Similar to the notation for population counts in the previous section, let $n_i(a)$ denote the number of nodes $u \in a$ with $y_u = i$ for arbitrary subsets $a \subset U$. Let $m_{ijkl}(a, b)$ be the count of pairs of nodes (u, v) such that $u \in a$, $v \in b$, $(y_u, y_v, x_{uv}, x_{vu}) = (ijkl)$, and $u < v$ if both u and v belong to $a \cap b$. An index replaced by a dot means summation over that index. For instance, according to the link-tracing designs described in section 3, only $m_{ijk\bullet}(s_0, s_1)$ is observed, not $m_{ijkl}(s_0, s_1)$.

With data from any of the link-tracing designs described in section 3, the likelihood function is

$$L(\theta, \lambda; d) = P(s | y_s, x_{s_0U}) \left(\prod_i \theta_i^{n_i(s)} \right) \left(\prod_{ijkl} \lambda_{ijkl}^{m_{ijk\bullet}(s_0, s_0)} \right) \times \left(\prod_{ijk} \lambda_{ijk\bullet}^{m_{ijk\bullet}(s_0, s_1)} \right) \prod_{v \in \bar{s}} \left[\sum_j \theta_j \prod_{ik} \lambda_{ijk\bullet}^{m_{ijk\bullet}(s_0, v)} \right]. \quad (4)$$

For the link-tracing designs in which all links, rather than a subsample, from the initial sample are traced, all of the elements in the submatrix $x_{s_0\bar{s}}$ are zero and $m_{i\bullet 0\bullet}(s_0, v) = n_i(s_0)$ for $v \in \bar{s}$, which simplifies the likelihood function to

$$L(\theta, \lambda; d) = P(s | y_s, x_{s_0U}) \left(\prod_i \theta_i^{n_i(s)} \right) \left(\prod_{ijkl} \lambda_{ijkl}^{m_{ijk\bullet}(s_0, s_0)} \right) \times \left(\prod_{ijk} \lambda_{ijk\bullet}^{m_{ijk\bullet}(s_0, s_1)} \right) \left[\sum_j \theta_j \prod_i \lambda_{ij0\bullet}^{n_i(s_0)} \right]^{n(\bar{s})}. \quad (5)$$

The factor $\prod_i \theta_i^{n_i(s)}$ gives the probability of the observed node values in the sample. The factor $\prod \lambda_{ijkl}^{m_{ijk\bullet}(s_0, s_0)}$ gives the probability of the observed dyad types within $s_0 \times s_0$ given the node values. The factor $\prod \lambda_{ijk\bullet}^{m_{ijk\bullet}(s_0, s_1)}$ gives the probability of the observed dyad types in $s_0 \times s_1$. Since x_{uv} but not x_{vu} is observed, for $u \in s_0$ and $v \in s_1$, the marginal probability that $x_{uv} = k$ given $y_u = i$ and $y_v = j$ is $\lambda_{ijk\bullet}$.

The final factor of (5), with square brackets, gives the probability that there are no arcs from the initial sample to \bar{s} . For a node v of the $n(\bar{s})$ nodes outside the sample, θ_j is the probability that $y_v = j$. From any of the $n_i(s_0)$ sample nodes $u \in s_0$ with $y_u = i$, the conditional probability of no link to v , that is, that $x_{uv} = 0$, $\lambda_{ij0\bullet}$.

More formally, the bracketed term can be obtained by conditioning on the number $n_j(\bar{s})$ of nodes of type j in \bar{s} . Conditional on $n_j(\bar{s})$, the probability that all the link indicators from s_0 to \bar{s} are zero is obtained as follows. From the $n_i(s_0)$ nodes of type i in s_0 to the $n_j(\bar{s})$ nodes of type j in \bar{s} , the probability that all links are zero is $\lambda_{ij0\bullet}^{n_i(s_0)n_j(\bar{s})}$. Using the binomial distribution of $n_j(\bar{s})$ with the law of total probability, the probability that all the links from s_0 to \bar{s} are zero, given y_s , is

$$\sum_{n_j(\bar{s})=0}^{n(\bar{s})} \binom{n(\bar{s})}{n_j(\bar{s})} \left(\prod_j \theta_j^{n_j(\bar{s})} \right) \left(\prod_{i,j} \lambda_{ij0\bullet}^{n_i(s_0)n_j(\bar{s})} \right) = \left[\sum_j \theta_j \prod_i \lambda_{ij0\bullet}^{n_i(s_0)} \right]^{n(\bar{s})}. \quad (6)$$

With the completed-wave design, the above likelihood expressions are simplified since the terms $m_{ijk\bullet}(s_0, s_1)$ are all zero, so that the factors involving these terms are all equal to one. We also note that $\lambda_{ij0\bullet} = 1 - \alpha_{ij}$ and $\lambda_{ji1\bullet} = \alpha_{ij}$ can be substituted to simplify the likelihood.

5.1.1 Estimative Likelihood Equations

The maximum likelihood estimators for the parameters θ_1 , α_{ij} , and β_k are obtained as the common solutions to the equations

$$\frac{d \log L}{d \theta_1} = \frac{d \log L}{d \alpha_{ij}} = \frac{d \log L}{d \beta_k} = 0 \quad (7)$$

for $i = 0, 1, j = 0, 1, k = 0, 2$. Differentiating the logarithm of the likelihood (5) with respect to θ_1 and setting equal to zero gives

$$\frac{d \log L}{d \theta_1} = \frac{\partial \log L}{\partial \theta_1} - \frac{\partial \log L}{\partial \theta_0} = 0$$

where the partial derivatives are given by

$$\frac{\partial \log L}{\partial \theta_k} = \frac{n_k(s)}{\theta_k} + n(\bar{s}) \frac{\prod_i \lambda_{ik0}^{n_i(s_0)}}{\sum_j \theta_j \prod_i \lambda_{ij0}^{n_i(s_0)}}$$

for $k = 0, 1$.

Moreover,

$$\frac{d \log L}{d \alpha_{ij}} = \frac{\partial \log L}{\partial \lambda_{ij10}} + \frac{\partial \log L}{\partial \lambda_{j01}} - \frac{\partial \log L}{\partial \lambda_{j00}} - \frac{\partial \log L}{\partial \lambda_{j100}} \quad (8)$$

and

$$\frac{d \log L}{d \beta_k} = \sum_{i+j=k} \left(\frac{\partial \log L}{\partial \lambda_{ij00}} + \frac{\partial \log L}{\partial \lambda_{ij11}} - \frac{\partial \log L}{\partial \lambda_{ij01}} - \frac{\partial \log L}{\partial \lambda_{ij10}} \right) \quad (9)$$

where the partial derivatives are given by

$$\begin{aligned} \frac{\partial \log L}{\partial \lambda_{ijkl}} &= \frac{m_{ijkl}(s_0, s_0)}{\lambda_{ijkl}} + \frac{m_{ijks}(s_0, s_1)}{\lambda_{ijks}} \\ &+ (1-k)n(\bar{s}) \frac{\theta_j n_i(s_0) \lambda_{ij0}^{n_i(s_0)-1}}{\sum_j \theta_j \prod_i \lambda_{ij0}^{n_i(s_0)}}. \end{aligned}$$

It is convenient to write the likelihood equation for θ_1 as

$$\frac{n_1(s)}{\theta_1} - \frac{n_0(s)}{\theta_0} + \frac{n(\bar{s})(\rho - 1)}{\theta_1 \rho + \theta_0} = 0 \quad (10)$$

where

$$\rho = \prod_{i=0}^1 \left(\frac{\lambda_{i10}}{\lambda_{i00}} \right)^{n_i(s_0)} = \prod_{i=0}^1 \left(\frac{1 - \alpha_{i1}}{1 - \alpha_{i0}} \right)^{n_i(s_0)}.$$

Note that $\rho = \rho_0^{n_0(s_0)} \rho_1^{n_1(s_0)}$, where $\rho_i = (1 - \alpha_{i1})/(1 - \alpha_{i0})$ is the ratio between the probabilities of no arc from an i -node to a positive and a zero node, respectively.

An interpretation of the influence of the graph structure on estimation of θ_1 is provided by considering the graph parameters α – and hence ρ – as fixed. Denote the sample proportion of positive nodes by $\hat{\theta}_c = n_1(s)/n(s)$. This is the conventional or naive estimator of θ_1 , using the sample proportion of positive nodes. If $\rho = 1$, then the maximum likelihood estimator $\hat{\theta}_1$ would be $\hat{\theta}_c$. If $\rho < 1$, then the maximum likelihood estimator $\hat{\theta}_1$ would be less than $\hat{\theta}_c$, and if $\rho > 1$, $\hat{\theta}_1 > \hat{\theta}_c$. In particular, $\alpha_{i1} = \alpha_{i0}$ for $i = 0, 1$ implies $\rho = 1$ and the maximum likelihood estimator is $\hat{\theta}_1 = \hat{\theta}_c$.

Consider for instance the case in which for any given value of y_u , a link from node u to node v is more likely when $y_v = 1$ than when $y_v = 0$, so that $\alpha_{i1} > \alpha_{i0}$, for $i = 0, 1$. Then $(1 - \alpha_{i1})/(1 - \alpha_{i0}) < 1$, for $i = 0, 1$, so that $\rho < 1$ and the maximum likelihood estimator $\hat{\theta}_1$ is less than the conventional estimator $\hat{\theta}_c$. One could say that the link-tracing design is leading investigators to an unrepresentatively high

proportion of positive nodes, and the maximum likelihood estimator is adjusting for this.

In specific cases some of the λ_{ijkl} might be set to zero and the likelihood equations have to be modified accordingly. Some specific cases will now be illustrated.

5.1.2 A Symmetric Model

Symmetric models have $\lambda_{ijkl} = 0$ for $k \neq l$ so that arcs are always mutual or, equivalently, they can be considered as undirected edges.

The full symmetric model has parameters $\lambda_{ijkk} = \lambda_{jikkk}$ for $i, j, k = 0, 1$, with $\lambda_{ij00} + \lambda_{ij11} = 1$. Here $\lambda_{ij11} = \beta_{i+j} = \alpha_{ij} = \alpha_{ji}$ and

$$\rho = \prod_{i=0}^1 \left(\frac{1 - \beta_{i+1}}{1 - \beta_i} \right)^{n_i(s_0)}.$$

Letting $m_{ijkl}(s_0, s) = r_{i+j, k+l}$, we obtain the maximum likelihood estimators as the solutions to the equations

$$\frac{n_1(s)}{\theta_1} - \frac{n_0(s)}{\theta_0} - \frac{n(\bar{s})(1 - \rho)}{\theta_0 + \rho \theta_1} = 0 \quad (11)$$

$$\frac{r_{02}}{\beta_0} - \frac{r_{00}}{1 - \beta_0} - \frac{n(\bar{s})n_0(s_0)\theta_0}{(1 - \beta_0)(\theta_0 + \rho \theta_1)} = 0 \quad (12)$$

$$\frac{r_{12}}{\beta_1} - \frac{r_{10}}{1 - \beta_1} - \frac{n(\bar{s})[n_1(s_0)\theta_0 + n_0(s_0)\rho \theta_1]}{(1 - \beta_1)(\theta_0 + \rho \theta_1)} = 0 \quad (13)$$

$$\frac{r_{22}}{\beta_2} - \frac{r_{20}}{1 - \beta_2} - \frac{n(\bar{s})n_1(s_0)\rho \theta_1}{(1 - \beta_2)(\theta_0 + \rho \theta_1)} = 0. \quad (14)$$

If the symmetric model is further simplified by assuming $\beta_0 = \beta_1 = 0$, there are only the two parameters θ_1 and β_2 , and the equations to be solved are

$$\theta_1 \beta_2 = r_{22}/N n_1(s_0)$$

and

$$\frac{N - n_1(s)/\theta_1}{N - n_0(s)/\theta_0} = (1 - \beta_2)^{n_1(s_0)}.$$

For instance suppose the value $y_u = 1$ indicates injection drug use and $x_{uv} = 1$ indicates u and v are injection partners, so that links are only possible between users and tracing these links can only add users to the sample. As an illustration, consider a population of size $N = 10,000$ with statistics $n_1(s_0) = 7$, $n_0(s_0) = 43$, $n_1(s) = 47$, and $r_{22} = 42$. The likelihood equations are $\theta_1 \beta_2 = 0.0006$ and $(10000 - 47/\theta_1)/(10000 - 43/\theta_0) = (1 - \beta_2)^7$, leading to the maximum likelihood estimators $\hat{\theta}_1 = 0.12$ and $\hat{\beta}_2 = 0.005$. The naive estimator for θ_1 in this case would be the sample proportion $47/90 = 0.52$ and the naive estimator for β_2 would be

$$42 / \binom{47}{2} = 0.039,$$

the proportion of links between users in the sample out of the number possible.

5.1.3 An Asymmetric Model

A specific asymmetric model has $\lambda_{ijkl} = \lambda_{ijk*} \lambda_{ij*l} = \lambda_{ijk*} \lambda_{jil*}$, so that all arcs are independent. Now $\beta_{i+j} = \alpha_{ij} \alpha_{ji}$ and we obtain the maximum likelihood estimators as the solutions to the equations

$$\rho = \frac{N - n_1(s)/\theta_1}{N - n_0(s)/\theta_0}$$

and

$$\frac{\alpha_{ij}}{1 - \alpha_{ij}} = \frac{m_{ij1}}{m_{j0} + n_i(s_0) \rho^j \theta_j (N - n_0(s)/\theta_0)}$$

for $i = 0, 1$ $j = 0, 1$, where $m_{ijk} = m_{ijk*}(s_0, s)$.

In particular, if we specify this asymmetric model by $\alpha_{ij} = i j \alpha$, so that arcs are possible with probability α only between marked nodes, then the equations to be solved are

$$\frac{N - n_1(s)/\theta_1}{N - n_0(s)/\theta_0} = (1 - \alpha)^{n_1(s_0)}$$

and

$$\frac{\alpha}{1 - \alpha} = \frac{m_{111}}{m_{110} + [N\theta_1 - n_1(s)]n_1(s_0)}.$$

Again, iterative methods are appropriate.

5.2 Predictive Likelihood for the Total of the Unobserved Node Values

For predicting the value of the unobserved random variable $n_1(\bar{s})$ from the observed data, the relevant likelihood is

$$L[\theta, \lambda; d, n_1(\bar{s})] = p(s | \mathbf{y}_s, \mathbf{x}_{s_0})$$

$$\begin{aligned} & \times \left(\prod_i \theta_i^{n_i(s) + n_i(\bar{s})} \right) \binom{n(\bar{s})}{N_1(\bar{s})} \left(\prod_{ijkl} \lambda_{ijkl}^{m_{ijkl}(s_0, s_0)} \right) \\ & \times \left(P \prod_{ijk} \lambda_{ijk*}^{m_{ijk*}(s_0, s_1)} \right) \left(\prod_j \lambda_{j0*}^{n_j(s_0) n_j(\bar{s})} \right). \end{aligned} \quad (15)$$

Use of the term "prediction" implies only that the object of inference is a random variable rather than a fixed, unknown parameter, and does not necessarily imply forecasting in time.

The estimative likelihood for $n_1(\bar{s})$ is obtained from (15) by substituting the estimates $\hat{\theta}$ and $\hat{\lambda}$ that maximize the (marginal) likelihood (5). The value of $n_1(\bar{s})$ maximizing the estimative likelihood would be the estimative maximum likelihood predictor of $n_1(\bar{s})$. While estimative likelihood methods tend to produce reasonable point predictions in

many cases, they are less useful as a basis for prediction intervals, since the estimates of the parameters are in essence treated as the true values (cf., Bjørnstad 1990, 1996, Lejeune and Faulkenberry 1982). For this reason, we emphasize the use of the profile predictive likelihood.

Rather than substituting fixed estimators of the parameters into (15) and maximizing this estimative likelihood with respect to $n_1(\bar{s})$, the likelihood (15) is now simultaneously maximized with respect to both parameters and $n_1(\bar{s})$. This means that for each value of $n_1(\bar{s})$ there are parameter values $\tilde{\theta}_j[n_1(\bar{s})]$ and $\tilde{\lambda}_{ijkl}[n_1(\bar{s})]$ which maximize (15) with respect to θ and λ . Substitution of these values into (15) defines the profile likelihood $L_p[n_1(\bar{s}); d]$ for $n_1(\bar{s})$. The value of $n_1(\bar{s})$ maximizing the profile likelihood is the profile maximum likelihood predictor of $n_1(\bar{s})$.

For any given value of $n_1(\bar{s})$, the likelihood is maximized where the derivatives with respect to the remaining parameters equal zero. The maximizing values of θ_j are straightforward and are given by

$$\tilde{\theta}_j = \frac{n_j(s) + n_j(\bar{s})}{N}. \quad (16)$$

For the remaining parameters we use $d \log L / d \alpha_{ij}$ and $d \log L / d \beta_k$ from (8) and (9), with the partial derivatives now given by

$$\frac{\partial \log L}{\partial \lambda_{ijkl}} = \frac{m_{ijkl}(s_0, s_0)}{\lambda_{ijkl}} + \frac{m_{ijk*}(s_0, s_1)}{\lambda_{ijk*}} + (1 - k) \frac{n_i(s_0) n_j(\bar{s})}{\lambda_{ij0*}} \quad (17)$$

Note that the $n_j(\bar{s})$ for $j = 0, 1$ are contained in (15) only in the factors

$$\binom{n(\bar{s})}{n_1(\bar{s})} \prod_j \Lambda_j^{n_j(\bar{s})}$$

where $\Lambda_j = \theta_j \prod_i \lambda_{ij0*}^{n_i(s_0)}$. Since L is proportional to a binomial probability with parameters $n(\bar{s})$ and $\Lambda_1 / (\Lambda_0 + \Lambda_1)$, it follows that the maximum of L over $n_1(\bar{s})$ is obtained for $n_1(\bar{s})$ equal to the integer closest to

$$\frac{n(\bar{s}) \Lambda_1}{\Lambda_0 + \Lambda_1} + \frac{\Lambda_1 - \Lambda_0}{2(\Lambda_0 + \Lambda_1)}$$

or either of the integers closest to this number if there are two of them. In fact (see, for instance, Feller 1957, p.140), the mode of a binomial distribution with parameters (n, p) is the integer in the interval $[(n+1)p - 1, (n+1)p]$ or either of the endpoints if they are integers. Thus, the mode is the integer or the integers that are closest to the interval midpoint $(n+1)p - (1/2) = np + (p-q)/2$, where $q = 1 - p$.

If initial values of the parameter estimators are obtained from the solution of (7) and substituted into the Λ_j , then a predicted value $n_1(\bar{s})$ is given as above. If this predicted value is inserted into (16) and (17), then new estimates of the parameters are obtained that can be substituted into the Λ_j to find a new predicted value of $n_1(\bar{s})$, continuing until the

values converge to the solution minimizing (15). Alternatively, the solution can be found by direct computation of the likelihood (15) for different values of $n_1(\bar{s})$, substituting the solutions obtained from (16) and (17) for the parameter values.

5.2.1 Example: Symmetric Model

The predictive likelihood equation (15) for the symmetric model is

$$L[\theta, \beta; d, n_1(\bar{s})] = P(s | \mathbf{y}_s, \mathbf{x}_{s_0 U}) \left(\prod_i \theta_i^{n_i(s) + n_i(\bar{s})} \right) \left(\frac{n(\bar{s})}{n_1(\bar{s})} \right) \times \left(\prod_{i,j} \beta_{i+j}^{m_{ij|11(s_0, s)} (1 - \beta_{i+j})^{m_{ij|00(s_0, s)} + n_i(s_0) n_j(\bar{s})}} \right). \quad (18)$$

Let $r_{kl} = r_{kl}(s_0, s)$ denote the count of node pairs in $s_0 \times s$ with total node value k and total number of links l . With the symmetric model, l can take only the values 0, indicating no link between the nodes, or 2, indicating a symmetric link. In particular, $r_{02} = m_{0011}(s_0, s)$, $r_{12} = m_{0111}(s_0, s) + m_{1011}(s_0, s)$, and $r_{22} = m_{1111}(s_0, s)$ denote the sample counts of links between nodes of total value k , for $k = 0, 1, 2$, respectively. With this notation the last factor in (18) can be written

$$\prod_{k=0}^2 \beta_k^{r_{k2}} (1 - \beta_k)^{r_{k0}} - \sum_{i,j} n_i(s_0) n_j(\bar{s}).$$

Denote by $c_k = c_k[n_1(\bar{s})]$ the number of possible node pairs in $s_0 \times U$ having total value k , so that

$$\begin{aligned} c_k &= r_{k*} + \sum_{i,j} n_i(s_0) n_j(\bar{s}) \\ &= \sum_{i,j} n_i(s_0) [n_j(s) + n_j(\bar{s})]. \end{aligned}$$

For any given value of $n_1(\bar{s})$, the likelihood is maximized by $\tilde{\theta}_i = [n_i(s) + n_i(\bar{s})]/N$ for $i = 0, 1$ and $\tilde{\beta}_k = r_{k2}/c_k$ for $k = 0, 1, 2$. Note that $\tilde{\theta}$ and the $\tilde{\beta}_k$ are functions of the unobserved variable $n_1(\bar{s})$.

The profile predictive likelihood function for $n_1(\bar{s})$ is obtained by substituting the maximizing values $\tilde{\theta}$ and $\tilde{\beta}_k$ for the parameters in (18), giving

$$\begin{aligned} L_p[n_1(\bar{s}); d] &= P(s | \mathbf{y}_s, \mathbf{x}_{s_0 U}) \left(\prod_i \left(\frac{n_i(s) + n_i(\bar{s})}{N} \right)^{n_i(s) + n_i(\bar{s})} \right) \\ &\times \left(\frac{n(\bar{s})}{n_1(\bar{s})} \right) \left(\prod_k \left(\frac{r_{k2}}{c_k} \right)^{r_{k2}} \left(1 - \frac{r_{k2}}{c_k} \right)^{c_k - r_{k2}} \right) \end{aligned}$$

which is a function of $n_1(\bar{s})$ alone. The maximum profile likelihood predictor of $n_1(\bar{s})$, easily obtained by straightforward computation, is an integer between 0 and $n(\bar{s})$ giving the largest value of (19).

5.3 On Assessing Accuracy of Estimates

For confidence intervals and other forms of inference, the inverse of the observed Fisher information $\mathbf{I}(\hat{\phi})$ is suggested, where $\hat{\phi}$ is the vector of parameter maximum likelihood estimates and \mathbf{I} is the matrix of negated second derivatives of the log likelihood function evaluated at those estimated values. The use of the observed, as opposed to expected, Fisher information to assess the accuracy of an estimate is described in Efron and Hinkley (1978). More recently, Lindsay and Li (1997) argue that the observed information gives a better assessment of the realized, as opposed to expected, error of the estimate. In developing large-sample approximations to the properties of the estimators of θ and λ it is important to make appropriate assumptions about how λ depends on N so that the graph model and the sample do not degenerate. See for instance the asymptotic results for some simple graph models given by Palmer (1985).

As with the calculation of the maximum likelihood estimates themselves, the calculation of the observed information matrix is not affected by the link-tracing sampling design, since the design is ignorable for likelihood based on inference. This is in contrast to the expected Fisher information, the value of which is affected by the design in addition to the graph model, unless the design is a conventional one not depending on any \mathbf{y} or \mathbf{x} values.

For a $(1 - \epsilon)$ -level prediction interval for a random variable such as $n_1(\bar{s})$, one method would be to use a central region having mass $(1 - \epsilon)$ of the normalized profile likelihood function for $n_1(\bar{s})$ (cf., Bjørnstad 1990, 1996). For the symmetric model, the $(1 - \epsilon)$ prediction interval for $n_1(\bar{s})$, is readily obtained by computing (19) for $n_1(\bar{s}) = 0, 1, 2, \dots$, until the computed values become negligible, normalizing by dividing by the cumulative total $\sum n_1(\bar{s}) = 0 L_p$, and using the $\epsilon/2$ and $1 - \epsilon/2$ quantiles as the interval endpoints.

ACKNOWLEDGEMENTS

Support for this research was provided by the National Science Foundation (DMS-9626102), the National Institutes of Health, National Institute on Drug Abuse (RO1 DA09872), and the Swedish Council for Research in the Humanities and the Social Sciences (HSFR F 0750/96).

REFERENCES

- BASU, D. (1969). Role of the sufficiency and likelihood principles in sample survey theory. *Sankhyā* A 31, 441-454.
- BIRNBAUM, Z.W., and SIRKEN, M.G. (1965). Design of sample surveys to estimate the prevalence of rare diseases: Three unbiased estimates. *Vital and Health Statistics*, 2, 11. Washington: Government Printing Office.
- BJØRNSTAD, J.F. (1990). Predictive likelihood: A review. *Statistical Science*, 5, 242-265.

- BJØRNSTAD, J.F. (1996). On the generalization of the likelihood function and the likelihood principle. *Journal of the American Statistical Association*, 91, 791-806.
- DAWID, A.P., and DICKEY, J.M. (1977). Likelihood and Bayesian inference from selectively reported data. *Journal of the American Statistical Association*, 72, 845-850.
- EFRON, B., and HINKLEY, D.V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information (with discussion). *Biometrika*, 65, 457-487.
- ERICKSON, B. (1979). Some problems of inference from chain data. *Sociological Methodology*, 10, 276-302.
- FIENBERG, S.E., MEYER, M.M., and WASSERMAN, S.S. (1985). Statistical analysis of multiple sociometric relations. *Journal of the American Statistical Association*, 80, 51-67.
- FLOURNOY, N., and ROSENBERGER, W.F., Eds. (1995). *Adaptive Designs*. Hayward, CA: Institute of Mathematical Statistics.
- FRANK, O. (1971). *Statistical Inference in Graphs*. Stockholm: Försvarets forskningsanstalt.
- FRANK, O. (1977a). Survey sampling in graphs. *Journal of Statistical Planning and Inference*, 1, 235-264.
- FRANK, O. (1977b). Estimation of graph totals. *Scandinavian Journal of Statistics*, 4, 81-89.
- FRANK, O. (1978a). Estimating the number of connected components in a graph by using a sampled subgraph. *Scandinavian Journal of Statistics*, 5, 177-188.
- FRANK, O. (1978b). Sampling and estimation in large social networks. *Social Networks*, 1, 91-101.
- FRANK, O. (1979a). Estimation of population totals by use of snowball samples. In *Perspectives on Social Network Research*, (Eds., P.W. Holland and S. Leinhardt). New York: Academic Press, 319-347.
- FRANK, O. (1979b). Moment properties of subgraph counts in stochastic graphs. *Annals of the New York Academy of Sciences*, 319, 207-218.
- FRANK, O. (1981). A survey of statistical methods for graph analysis. *Sociological Methodology*, 110-155.
- FRANK, O. (1988). Random sampling and social networks: a survey of various approaches. *Mathematiques, Informatique et Sciences humaines*, 26, 19-33.
- FRANK, O. (1997). Composition and structure of social networks. *Mathematiques, Informatique et Sciences humaines*, 35, 11-23.
- FRANK, O., and HARARY, F. (1982). Cluster inference by using transitivity indices in empirical graphs. *Journal of the American Statistical Association*, 77, 835-840.
- FRANK, O., and SNIJDERS, T. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*, 10, 53-67.
- FRANK, O., and STRAUSS, D. (1986). Markov graphs. *Journal of the American Statistical Association*, 81, 832-842.
- FRIEDMAN, S.R., NEAIGUS, A., JOSE, B., CURTIS, R., GOLDSTEIN, M., ILDEFONSO, G., ROTHENBERG, R.B., and DES JARLAIS, D.C. (1997). Sociometric risk networks and HIV risk. *American Journal of Public Health*. In press.
- GODAMBE, V.P. (1966). A new approach to sampling from finite populations. 1. *Journal of the Royal Statistical Society B*, 28, 310-319.
- GOODMAN, L.A. (1961). Snowball sampling. *Annals of Mathematical Statistics*, 32, 148-170.
- GRANOVETTER, M. (1976). Network sampling: some first steps. *American Journal of Sociology*, 81, 1287-1303.
- HOLLAND, P.W., LASKEY, K.B., and LEINHARDT, S. (1983). Stochastic block-models: First steps. *Social Networks*, 5, 109-137.
- HOLLAND, P.W., and LEINHARDT, S. (1981). An exponential family of probability distributions for directed graphs (with discussion). *Journal of the American Statistical Association*, 76, 33-65.
- JANSSON, I. (1997). On statistical modeling of social networks. Ph.D. Thesis, Stockholm University.
- KALTON, G., and ANDERSON, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society A*, 149, 65-82.
- KARLBERG, M. (1997). Triad count estimation and transitivity testing in graphs and digraphs. Ph.D. Thesis, Stockholm University.
- KLOVDAHL, A.S. (1989). Urban social networks: Some methodological problems and possibilities. In *The Small World*, (Ed. M. Kochen). Norwood, NJ: Ablex Publishing, 176-210.
- LEJEUNE, M., and FAULKENBERRY, G.D. (1982). A simple predictive density function. *Journal of the American Statistical Association*, 77, 654-657.
- LEVY, P.S. (1977). Optimum allocation in stratified random network sampling for estimating the prevalence of attributes in rare populations. *Journal of the American Statistical Association*, 72, 758-763.
- LEVY, P.S., and LEMESHOW, S. (1991). *Sampling of Populations: Methods and Applications*. New York: Wiley.
- LINDSAY, B.G., and LI, B. (1997). On second-order optimality of the observed Fisher information. *Annals of Statistics*, 25, 2172-2199.
- MORGAN, D.L., and RYTINA, S. (1977). Comment on "Network sampling: some first steps" by Mark Granovetter. *American Journal of Sociology*, 83, 722-727.
- NEAIGUS, A., FRIEDMAN, S.R., GOLDSTEIN, M.F., ILDEFONSO, G., CURTIS, R., and JOSE, B. (1995). Using dyadic data for a network analysis of HIV infection and risk behaviors among injection drug users. In (Eds., R.H. Needle, S.G. Genser, and R.T. Trotter II) *Social Networks, Drug Abuse, and HIV Transmission*. NIDA Research Monograph 151. Rockville, MD: National Institute of Drug Abuse, 20-37.
- NEAIGUS, A., FRIEDMAN, S.R., JOSE, B., GOLDSTEIN, M.F., CURTIS, R., ILDEFONSO, G., and DES JARLAIS, D.C. (1996). High-risk personal networks and syringe sharing as risk factors for HIV infection among new drug injectors. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, 11, 499-509.
- PALMER, E.M. (1985). *Graphical Evolution*. New York: Wiley.

- ROBINS, G.L. (1998). Personal attributes in inter-personal contexts: statistical models for individual characteristics and social relationships. Ph.D. Thesis, University of Melbourne.
- ROSENBERGER, W.F. (1996). New directions in adaptive designs. *Statistical Science*, 11, 137-149.
- ROTHENBERG, R.B., WOODHOUSE, D.E., POTTERAT, J.J., MUTH, S.Q., DARROW, W.W., and KLOVDAHL, A.S. (1995). Social networks in disease transmission: The Colorado Springs study. In (Eds., R.H. Needle, S.G. Genser, and R.T. Trotter II), *Social Networks, Drug Abuse, and HIV Transmission*. NIDA Research Monograph 151. Rockville, MD: National Institute of Drug Abuse, 3-19.
- RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- SCOTT, A.J. (1977). On the problem of randomization in survey sampling. *Sankhyā C*, 39, 1-9.
- SCOTT, A.J., and SMITH, T.M.F. (1973). Survey design, symmetry, and posterior distributions. *Journal of the Royal Statistical Society B*, 35, 57-60.
- SIRKEN, M.G. (1970). Household surveys with multiplicity. *Journal of the American Statistical Association*, 63, 257-266.
- SIRKEN, M.G. (1972a). Stratified sample surveys with multiplicity. *Journal of the American Statistical Association*, 67, 224-227.
- SIRKEN, M.G. (1972b). Variance components of multiplicity estimators. *Biometrics*, 28, 869-873.
- SIRKEN, M.G., and LEVY, P.S. (1974). Multiplicity estimation of proportions based on ratios of random variables. *Journal of the American Statistical Association*, 69, 68-73.
- SNIJDERS, T.A.B. (1992). Estimation on the basis of snowball samples: how to weight. *Bulletin de Methodologie Sociologique*, 36, 59-70.
- SPREEN, M. (1992). Rare populations, hidden populations, and link-tracing designs; what and why? *Bulletin de Methodologie Sociologique*, 36, 34-58.
- SPREEN, M. (1998). Sampling personal network structures: statistical inference in ego-graphs. Ph.D. Thesis, University of Groningen.
- SPREEN, M., and ZWAAGSTRA, R. (1994). Personal network sampling, outdegree analysis and multilevel analysis: introducing the network concept in studies of hidden populations. *International Sociology*, 9, 475-491.
- SUDMAN, S., SIRKEN, M.G., and COWAN, C.D. (1988). Sampling rare and elusive populations. *Science*, 240, 991-996.
- SUGDEN, R.A., and SMITH, T.M.F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, 71, 495-506.
- THOMPSON, S.K. (1990). Adaptive cluster sampling. *Journal of the American Statistical Association*, 85, 1050-1059.
- THOMPSON, S.K. (1997). Adaptive sampling in behavioral surveys. In (Eds., L. Harrison, and A. Hughes), *The Validity of Self-Reported Drug Use: Improving the Accuracy of Survey Estimates*. NIDA Research Monograph 167, Rockville, MD: National Institute of Drug Abuse, 296-319.
- THOMPSON, S.K., and SEBER, G.A.F. (1996). *Adaptive Sampling*. New York: Wiley.
- van METER, K.M. (1990). Methodological and design issues: techniques for assessing the representatives of snowball samples. In (Ed., E.Y. Lambert), *The Collection and Interpretation of Data from Hidden Populations*. NIDA Monograph 98. Rockville, MD: National Institute on Drug Abuse, 31-43.
- WANG, Y.J., and WONG, G.Y. (1987). Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82, 8-19.
- WASSERMAN, S. (1980). Analyzing social networks as stochastic processes. *Journal of the American Statistical Association*, 75, 280-294.
- WASSERMAN, S., and FAUST, K. (1994). *Social Network Analysis: Methods and Applications*. New York: Cambridge University Press.
- WATTERS, J.K., and BIERNACKI, P. (1989). Targeted sampling: Options for the study of hidden populations. *Social Problems*, 36, 416-430.
- WEI, L.J., SMYTHE, R.T., LIN, D.Y., and PARK, T.S. (1990). Statistical inference with data-dependent treatment allocation rules. *Journal of the American Statistical Association*, 85, 156-162.
- WELLMAN, B., FRANK, O., ESPINOZA, V., LUNDQUIST, S., and WILSON, C. (1991). Integrating individual, relational and structural analysis. *Social Networks*, 13, 223-249.

Calibration and Restricted Weights

ALAIN THÉBERGE¹

ABSTRACT

To better understand the impact of imposing a restricted region on calibration weights, the author reviews the latter's asymptotic behaviour. Necessary and sufficient conditions are provided for the existence of a solution to the calibration equation with weights within given intervals. A more general formulation of the calibration problem leads to a compromise between the need to satisfy the calibration equation and the attempt to obtain weights that are close to Horvitz-Thompson weights. If the requirements for the calibration equation are relaxed, then various estimation methods with restricted weights can be used. The estimators that are introduced usually have the same asymptotic properties as the calibration estimator with no weight restrictions, and some have weights which can be calculated explicitly, without any iterative process. The author shows how these estimators can be adapted to take advantage of a synthetic estimator. An approach similar to that used to restrict weights is applied to outliers.

KEY WORDS: Small domains; Moore-Penrose inverse; Inequality solutions; Asymptotic properties; Outliers.

1. INTRODUCTION

The calibration estimator has good asymptotic properties. However, for samples of small size, or if calibration is done at the domain level and some of the domains involve few observations, the weights of such an estimator can include extreme values. One way of overcoming this problem consists in using the calibration method with distance measurements which restrict the weights of observations to certain intervals about the sampling weights. This approach was developed by Deville and Särndal (1992). Other methods aimed at providing robust estimates satisfying the calibration equation can be found in Duchesne (1999). That paper contains an extensive bibliography on robust estimators. However, there is no guaranteed solution to the calibration equation with restricted weights. Even when such weights exist, the statistician might prefer solving the problem of extreme weights by relaxing somewhat the requirements for the calibration equation, instead of tightening the constraints on the weights by using a distance measurement that is more "restrictive". This paper provides a formulation of the calibration problem which offers more flexibility to the statistician. The problem in fact is one of minimization similar to that encountered in ridge regression. Bardsley and Chambers (1984) encountered this same minimization problem in their search for model-based estimators. This formulation of the calibration problem can be used to restrict weights without the use of special distances between calibrated weights and Horvitz-Thompson weights. Rao and Singh (1997) combined this approach with iterative methods using distance measurements. Other ways of restricting weights will also be reviewed.

In the next section, the calibration method is outlined without applying limits to the values of weights. The

calibration problem thus outlined does not assume there is a solution to the calibration equation. The asymptotic properties of calibrated weights are discussed. These properties have a bearing on the asymptotic behaviour of the estimators whose weights are restricted. In section 3, necessary and sufficient conditions are provided for the existence of restricted weights which satisfy the calibration equation. Section 4 discusses how the estimation problem can be formulated by varying the importance attributed to the calibration equation. Section 5 provides various means of restricting weights without recourse to a specific distance. Section 6 introduces an estimator with restricted weights which is useful for small domains. Finally, in section 7, outliers are discussed in terms of a method similar to that used to deal with extreme weights.

2. CALIBRATION

Let $Y \in \mathbb{R}^{N \times d}$ denote a matrix of d variables of interest for a population of size N , and let $c \in \mathbb{R}^N$ denote a vector of known constants; a sample s of size n is drawn, and the subscript s is used to designate the sub-vectors or sub-matrices corresponding to the sample. We wish to estimate $Y'c$ using $Y'_s w_s$, where $w_s \in \mathbb{R}^n$ is a weight vector for the sampled units. For a vector v and a positive diagonal matrix F of identical dimension, we define $\|v\|_F^2 = v'Fv$. For an auxiliary information matrix $X \in \mathbb{R}^{N \times p}$, $A \in \mathbb{R}^{N \times N}$ the diagonal matrix of sampling weights, given positive diagonal matrices $U_s \in \mathbb{R}^{n \times n}$ and $T \in \mathbb{R}^{p \times p}$, we seek, among the weight vectors $w_s \in \mathbb{R}^n$ which minimize $\|X'_s w_s - X'c\|_T^2$, the one which minimizes $D_s(w_s) = \|w_s - A_s c_s\|_{U_s}^2$. This formulation of the problem, which does not assume the existence of weights satisfying the calibration equation, $X'_s w_s = X'c$, can be found in Théberge (1999). The

¹ Alain Théberge, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 Canada.

solution represents the vector of calibrated weights \mathbf{w}_{cal} . We have

$$\mathbf{w}_{\text{cal}} = \mathbf{A}_s \mathbf{c}_s + \mathbf{U}_s^{-1} \mathbf{X}_s' \mathbf{T}^{1/2} (\mathbf{T}^{1/2} \mathbf{X}_s' \mathbf{U}_s^{-1} \mathbf{X}_s \mathbf{T}^{1/2})^\dagger \mathbf{T}^{1/2} (\mathbf{X}' \mathbf{c} - \mathbf{X}_s' \mathbf{A}_s \mathbf{c}_s), \quad (1)$$

where \mathbf{F}^\dagger denotes the Moore-Penrose inverse of the matrix \mathbf{F} .

To better review the asymptotic properties of calibration estimators with restricted weights, let us now examine the behaviour of \mathbf{w}_{cal} when $n \rightarrow \infty$. We assume there exists an asymptotic setup in which the size of the population and the size of the sample tend towards infinity (see for example Isaki and Fuller (1982)), and for which we have

$$\mathbf{Y}' \mathbf{c} = O_p(N^\gamma) \quad (\gamma \geq 0)$$

$$\mathbf{X}' \mathbf{c} - \mathbf{X}_s' \mathbf{A}_s \mathbf{c}_s = O_p(n^{-1/2} N^\gamma) \quad (2)$$

$$\mathbf{T}^{1/2} \mathbf{X}_s' \mathbf{U}_s^{-1} \mathbf{X}_s \mathbf{T}^{1/2} = O_p(n).$$

It follows that $(\mathbf{T}^{1/2} \mathbf{X}_s' \mathbf{U}_s^{-1} \mathbf{X}_s \mathbf{T}^{1/2})^\dagger = O_p(n^{-1})$, since one of the properties of the Moore-Penrose inverse of a matrix \mathbf{F} is $\mathbf{F}^\dagger \mathbf{F} \mathbf{F}^\dagger = \mathbf{F}^\dagger$. Usually, we can expect to have $\gamma = 1$ when each element of the vector \mathbf{c} has a value of 1 (estimate of a total), and $\gamma = 0$ when each element of \mathbf{c} has a value of $1/N$ (estimate of a mean). For conditions (2) we therefore have,

$$\begin{aligned} \mathbf{w}_{\text{cal}} - \mathbf{A}_s \mathbf{c}_s &= \mathbf{U}_s^{-1} \mathbf{X}_s' \mathbf{T}^{1/2} (\mathbf{T}^{1/2} \mathbf{X}_s' \mathbf{U}_s^{-1} \mathbf{X}_s \mathbf{T}^{1/2})^\dagger \mathbf{T}^{1/2} (\mathbf{X}' \mathbf{c} - \mathbf{X}_s' \mathbf{A}_s \mathbf{c}_s) \\ &= O_p(n^{-1}) O_p(n^{-1/2} N^\gamma) \\ &= O_p(n^{-3/2} N^\gamma). \end{aligned} \quad (3)$$

Thus $\mathbf{w}_{\text{cal}} - \mathbf{A}_s \mathbf{c}_s$ converges in probability to $\mathbf{0}$, if

$$\lim_{n, N \rightarrow \infty} n^{-3/2} N^\gamma = 0.$$

For an asymptotic setup such as that of Brewer (1979) in which the sampling fraction n/N is constant, or any setup for which the sampling fraction converges to a positive number, this condition is verified if $\gamma < 3/2$.

Writing $\mathbf{w}_{\text{cal}} = \mathbf{A}_s \mathbf{c}_s + \mathbf{U}_s^{-1} \mathbf{X}_s' \mathbf{T}^{1/2} \mathbf{H}_s^\dagger \mathbf{T}^{1/2} (\mathbf{X}' \mathbf{c} - \mathbf{X}_s' \mathbf{A}_s \mathbf{c}_s)$, where $\mathbf{H}_s = \mathbf{T}^{1/2} \mathbf{X}_s' \mathbf{U}_s^{-1} \mathbf{X}_s \mathbf{T}^{1/2}$, we have

$$\begin{aligned} D_s(\mathbf{w}_{\text{cal}}) &= (\mathbf{X}' \mathbf{c} - \mathbf{X}_s' \mathbf{A}_s \mathbf{c}_s)' \mathbf{T}^{1/2} \mathbf{H}_s^\dagger \mathbf{H}_s \mathbf{H}_s^\dagger \mathbf{T}^{1/2} (\mathbf{X}' \mathbf{c} - \mathbf{X}_s' \mathbf{A}_s \mathbf{c}_s) \\ &= (\mathbf{X}' \mathbf{c} - \mathbf{X}_s' \mathbf{A}_s \mathbf{c}_s)' \mathbf{T}^{1/2} \mathbf{H}_s^\dagger \mathbf{T}^{1/2} (\mathbf{X}' \mathbf{c} - \mathbf{X}_s' \mathbf{A}_s \mathbf{c}_s) \\ &= O_p(n^{-1/2} N^\gamma) O_p(n^{-1}) O_p(n^{-1/2} N^\gamma) \\ &= O_p(n^{-2} N^{2\gamma}). \end{aligned} \quad (4)$$

Again for an asymptotic setup in which the sampling fraction converges to a positive number, we have $D_s(\mathbf{w}_{\text{cal}})$ converging in probability to 0, if $\gamma < 1$. Thus there are cases, e.g. for the estimate of a total, where $\mathbf{w}_{\text{cal}} - \mathbf{A}_s \mathbf{c}_s$ converges in probability to $\mathbf{0}$, but where $D_s(\mathbf{w}_{\text{cal}}) = \|\mathbf{w}_{\text{cal}} - \mathbf{A}_s \mathbf{c}_s\|_{\mathbf{U}_s}^2$ does not converge to 0.

3. CALIBRATION EQUATION SOLUTIONS AND RESTRICTED WEIGHTS

Even in the absence of weight restrictions, there might not be a solution to the calibration equation. By applying Graybill (1983, 113) to the calibration problem, we find that the calibration equation $\mathbf{X}_s' \mathbf{w}_s = \mathbf{X}' \mathbf{c}$ can be solved if and only if $(\mathbf{X}_s' \mathbf{X}_s)^+ \mathbf{X}_s' \mathbf{c} = \mathbf{X}' \mathbf{c}$. If there is a solution, the calibrated weights might be negative or exceptionally large. Deville and Särndal (1992) proposed using various distance measures other than a weighted sum of squares to measure the distance between Horvitz-Thompson weights and calibrated weights, so as to restrict the weights to certain intervals while satisfying the calibration equation. This approach can only work if there are in these intervals weights which satisfy the calibration equation. The main goal of this section is to find necessary and sufficient conditions for the existence of a weight vector \mathbf{w}_s within given bounds, such that the estimates of totals for auxiliary variables are also bounded. In other words, we are seeking conditions for the existence of a vector \mathbf{w}_s such that $\mathbf{w}_s^{(L)} \leq \mathbf{w}_s \leq \mathbf{w}_s^{(H)}$ and $\mathbf{t}^{(L)} \leq \mathbf{X}_s' \mathbf{w}_s \leq \mathbf{t}^{(H)}$, where $\mathbf{w}_s^{(L)}$, $\mathbf{w}_s^{(H)}$, $\mathbf{t}^{(L)}$ and $\mathbf{t}^{(H)}$ are given. In particular, by assuming $\mathbf{t}^{(L)} = \mathbf{t}^{(H)} = \mathbf{X}' \mathbf{c}$, we would obtain conditions for the existence of weights restricted to the intervals $\mathbf{w}_s^{(L)} \leq \mathbf{w}_s \leq \mathbf{w}_s^{(H)}$, satisfying the calibration equation.

A first step is provided by the following Fan (1956) theorem. It is formulated here for a matrix \mathbf{M} of finite dimension, although the proof provided by Fan also applies to a matrix of infinite dimension. The theorem uses the kernel of \mathbf{M}' , $N(\mathbf{M}')$, defined as the set of vectors $\boldsymbol{\alpha}$ such that $\mathbf{M}' \boldsymbol{\alpha} = \mathbf{0}$.

Theorem: Let $\mathbf{M} \in \mathbb{R}^{m \times n}$ and $\mathbf{l} \in \mathbb{R}^m$, $\exists \mathbf{w} \in \mathbb{R}^n$ such that $\mathbf{M} \mathbf{w} \geq \mathbf{l}$ if and only if for any $\boldsymbol{\lambda} \geq \mathbf{0}$ in $N(\mathbf{M}')$, we have $\mathbf{l}' \boldsymbol{\lambda} \leq 0$.

Corollary: Let $\mathbf{M} \in \mathbb{R}^{m \times n}$ and $\mathbf{l}, \mathbf{h} \in \mathbb{R}^m$, $\exists \mathbf{w} \in \mathbb{R}^n$ such that $\mathbf{l} \leq \mathbf{M} \mathbf{w} \leq \mathbf{h}$ if and only if first $\mathbf{l} \leq \mathbf{h}$ and secondly $\boldsymbol{\lambda} \in N(\mathbf{M}') \Rightarrow -\mathbf{l}' \boldsymbol{\lambda}_- \leq \mathbf{h}' \boldsymbol{\lambda}_+$, where $\boldsymbol{\lambda}_+ = \max(\boldsymbol{\lambda}, \mathbf{0})$ and $\boldsymbol{\lambda}_- = \min(\boldsymbol{\lambda}, \mathbf{0})$ with the extrema taken elementwise.

The corollary is obtained by using the theorem with

$$\mathbf{M} = \begin{pmatrix} \mathbf{M} \\ -\mathbf{M} \end{pmatrix}, \mathbf{l} = \begin{pmatrix} \mathbf{l} \\ -\mathbf{h} \end{pmatrix} \text{ and } \boldsymbol{\lambda} = \begin{pmatrix} -\boldsymbol{\lambda}_- \\ \boldsymbol{\lambda}_+ \end{pmatrix}$$

Let p denote the dimension of $N(\mathbf{M}')$. If p is equal to zero, then $\boldsymbol{\lambda} \in N(\mathbf{M}')$ implies $\boldsymbol{\lambda} = \mathbf{0}$, and the condition of the theorem (or the similar condition of the corollary) is obviously met. If p is equal to one, then $\boldsymbol{\lambda} \in N(\mathbf{M}')$ implies that $\boldsymbol{\lambda}$ is a multiple of a vector \mathbf{z} , and it is sufficient to

check the condition for $\lambda = z$ and $\lambda = -z$. If we use the property $(-\lambda)_- = -(\lambda)_+$, the problem outlined at the beginning of the section can now be resolved if X_s is a vector. The corollary with

$$M = \begin{pmatrix} I_n \\ X_s' \end{pmatrix}, l = \begin{pmatrix} w^{(L)} \\ t^{(L)} \end{pmatrix}, h = \begin{pmatrix} w^{(H)} \\ t^{(H)} \end{pmatrix},$$

and the fact that

$$z = \begin{pmatrix} -X_s \\ 1 \end{pmatrix}$$

spans $N(M')$, provide the necessary and sufficient conditions

$$\begin{aligned} w^{(L)} &\leq w^{(H)} \\ t^{(L)} &\leq t^{(H)} \\ (X_s)_+ w^{(L)} + (X_s)_- w^{(H)} &\leq t^{(H)} \\ t^{(L)} &\leq (X_s)_+ w^{(H)} + (X_s)_- w^{(L)}. \end{aligned} \quad (5)$$

The third inequality in (5) states that the weighted total of the auxiliary variable must not be greater than $t^{(H)}$, when the smallest possible weight $w^{(L)}$ is given to units for which the auxiliary variable is positive, and when the greatest possible weight $w^{(H)}$ is given to units for which the auxiliary variable is negative. The fourth inequality in (5) states that the weighted total of the auxiliary variable must not be less than $t^{(L)}$, when the largest possible weight is given to units for which the auxiliary variable is positive, and when the smallest possible weight is given to units for which the auxiliary variable is negative.

Even for $p > 1$, it is sufficient to check the condition of the corollary for a finite number of values of λ . Let $V \in \mathbb{R}^{m \times p}$, $2 \leq p \leq m$ denote a matrix whose columns form a basis for $N(M')$. It is always possible to construct V such that p of its rows, v_1, v_2, \dots, v_m , are the unit vectors of \mathbb{R}^p , and we will assume that V is of this form. It will be shown in Appendix A that it is sufficient to check the condition of the corollary for vectors $\lambda = V\phi$ and $\lambda = -V\phi$, where $\phi = (\phi_1, \dots, \phi_p)'$ is a non-zero vector satisfying $v_i' \phi = 0$ for a subset of $(p-1)$ linearly independent vectors v_i . We must therefore check the condition at the most for $\binom{m}{p-1}$ vectors ϕ , i.e. at the most $2 \binom{m}{p-1}$ values of λ .

Using the corollary with

$$M = \begin{pmatrix} I_n \\ X_s' \end{pmatrix}, l = \begin{pmatrix} w^{(L)} \\ t^{(L)} \end{pmatrix}, h = \begin{pmatrix} w^{(H)} \\ t^{(H)} \end{pmatrix},$$

and noting that the columns of

$$V = \begin{pmatrix} -X_s \\ I_p \end{pmatrix}$$

form a basis for $N(M')$, we obtain the following necessary and sufficient conditions for the existence of a solution to

the problem mentioned at the beginning of this section whenever $X_s \in \mathbb{R}^{n \times p}$ with $p > 1$. We must have $w^{(L)} \leq w^{(H)}$, $t^{(L)} \leq t^{(H)}$, and for each subset of $(p-1)$ linearly independent rows of

$$V = \begin{pmatrix} -X_s \\ I_p \end{pmatrix}$$

it is necessary that

$$\begin{aligned} (X_s \phi)' w^{(L)} - \phi' t^{(L)} &\leq -(X_s \phi)' w^{(H)} + \phi' t^{(H)} \\ -(X_s \phi)' w^{(L)} + \phi' t^{(L)} &\leq (X_s \phi)' w^{(H)} - \phi' t^{(H)} \end{aligned} \quad (6)$$

for a non-zero vector $\phi \in \mathbb{R}^p$ orthogonal to each row of the subset. The second inequality in (6) is obtained from the first by changing the sign of ϕ .

If $V_{\text{sub}} \in \mathbb{R}^{p \times p}$ is a non-singular matrix whose rows are rows of V , then each column of V_{sub}^{-1} is a vector perpendicular to $(p-1)$ linearly independent rows of V . Hence the following result:

There exists a weight vector w_s such that $w^{(L)} \leq w_s \leq w^{(H)}$ and $t^{(L)} \leq X_s' w_s \leq t^{(H)}$ if and only if $w^{(L)} \leq w^{(H)}$, $t^{(L)} \leq t^{(H)}$ and

$$\begin{aligned} (X_s V_{\text{sub}}^{-1})' w^{(L)} - (V_{\text{sub}}^{-1})' t^{(L)} &\leq -(X_s V_{\text{sub}}^{-1})' w^{(H)} \\ &\quad + (V_{\text{sub}}^{-1})' t^{(H)} \\ -(X_s V_{\text{sub}}^{-1})' w^{(L)} + (V_{\text{sub}}^{-1})' t^{(L)} &\leq (X_s V_{\text{sub}}^{-1})' w^{(H)} \\ &\quad - (V_{\text{sub}}^{-1})' t^{(H)} \end{aligned} \quad (7)$$

for all non-singular matrixes $V_{\text{sub}} \in \mathbb{R}^{p \times p}$ whose rows are rows of

$$V = \begin{pmatrix} -X_s \\ I_p \end{pmatrix}.$$

These conditions are somewhat redundant. For example, if inequalities (7) are met for $V_{\text{sub}} = V_1$, then they are necessarily met for any matrix V_2 obtained from V_1 through a permutation of rows.

Another example is provided by weighting observations in a contingency table. Assuming $\hat{N}_{ij} = n_{ij} w_{ij}$ ($i = 1, 2, \dots, R; j = 1, 2, \dots, C$), where n_{ij} is the number of observations in cell (i, j) of a contingency table and w_{ij} is the weight of each of these observations, we wish to know if there are weights w_{ij} such that \hat{N}_{ij} satisfies certain constraints. For example, motivated by the problem of convergence of the raking ratio procedure, Bacharach (1965) provided necessary and sufficient conditions for the existence of weights w_{ij} such that $\hat{N}_{ij} \geq 0$, $\sum_{i=1}^R \hat{N}_{ij} = N_j$ ($j = 1, \dots, C$), $\sum_{j=1}^C \hat{N}_{ij} = N_i$ ($i = 1, \dots, R$), where the values of N_j and N_i are given. The following result, demonstrated in Appendix B, is more general.

For arbitrary constants $N_{ij}^{(L)}, N_{ij}^{(H)}, N_j^{(L)}, N_j^{(H)}, N_{i..}^{(L)}, N_{i..}^{(H)}, N_{..}^{(L)}$, and $N_{..}^{(H)}$, there are \hat{N}_{ij} such that

$$N_{ij}^{(L)} \leq \hat{N}_{ij} \leq N_{ij}^{(H)} \quad i=1, \dots, R; j=1, \dots, C;$$

$$N_j^{(L)} \leq \sum_{i=1}^R \hat{N}_{ij} \leq N_j^{(H)} \quad j=1, \dots, C;$$

$$N_{i..}^{(L)} \leq \sum_{j=1}^C \hat{N}_{ij} \leq N_{i..}^{(H)} \quad i=1, \dots, R;$$

$$N_{..}^{(L)} \leq \sum_{i=1}^R \sum_{j=1}^C \hat{N}_{ij} \leq N_{..}^{(H)},$$

if and only if

$$N_{ij}^{(L)} \leq N_{ij}^{(H)} \quad i=1, \dots, R; j=1, \dots, C;$$

$$N_j^{(L)} \leq N_j^{(H)} \quad j=1, \dots, C;$$

$$N_{i..}^{(L)} \leq N_{i..}^{(H)} \quad i=1, \dots, R;$$

$$N_{..}^{(L)} \leq N_{..}^{(H)}$$

$$\begin{aligned} & \sum_{j \in T} \left(N_j^{(L)} - \sum_{i \in S} N_{ij}^{(H)} \right) \\ & \leq \sum_{i \in S} \left(N_{i..}^{(H)} - \sum_{j \notin T} N_{ij}^{(L)} \right) \\ & \sum_{i \in S} \left(N_{i..}^{(L)} - \sum_{j \notin T} N_{ij}^{(H)} \right) \\ & \leq \sum_{j \in T} \left(N_j^{(H)} - \sum_{i \notin S} N_{ij}^{(L)} \right) \\ & N_{..}^{(L)} + \sum_{j \notin T} \left(N_j^{(H)} - \sum_{i \in S} N_{ij}^{(H)} \right) \\ & \leq \sum_{i \in S} \left(N_{i..}^{(H)} - \sum_{j \in T} N_{ij}^{(L)} \right) + \sum_{j=1}^J N_j^{(H)} \\ & \sum_{i=1}^I N_{i..}^{(L)} + \sum_{j \notin T} \left(N_j^{(L)} - \sum_{i \in S} N_{ij}^{(H)} \right) \\ & \leq \sum_{i \in S} \left(N_{i..}^{(L)} - \sum_{j \in T} N_{ij}^{(L)} \right) + N_{..}^{(H)} \end{aligned} \quad (9)$$

for any $S \subseteq \{1, 2, \dots, R\}$, $T \subseteq \{1, 2, \dots, C\}$.

The number of inequalities to be checked can be reduced. For example, instead of checking

$$\sum_{j \in T} \left(N_j^{(L)} - \sum_{i \in S} N_{ij}^{(H)} \right) \leq \sum_{i \in S} \left(N_{i..}^{(H)} - \sum_{j \notin T} N_{ij}^{(L)} \right)$$

for any $S \subseteq \{1, 2, \dots, R\}$, and $T \subseteq \{1, 2, \dots, C\}$, it can be readily shown that an equivalent procedure would be to check that

$$\sum_{j \in T} N_j^{(L)} \leq \sum_{i=1}^R \min \left(\left(N_{i..}^{(H)} - \sum_{j \notin T} N_{ij}^{(L)} \right), \sum_{j \in T} N_{ij}^{(H)} \right)$$

for any $T \subseteq \{1, 2, \dots, C\}$.

4. MITIGATED CALIBRATION

There may be dissatisfaction with the two-step approach of calibration, where an attempt is first made to find weight vectors that best satisfy the calibration equation, and then from this set of vectors to find the one which comes closest to Horvitz-Thompson weights. For small samples, this method may lead to weights which the statistician will find too far from Horvitz-Thompson weights.

There may be a preference for varying the importance attributed to the calibration equation relative to the norm of $\mathbf{w}_s - \mathbf{A}_s \mathbf{c}_s$. Thus, there may be a desire to find a weight vector \mathbf{w}_s which minimizes

$$\left\| \begin{pmatrix} \mathbf{w}_s - \mathbf{A}_s \mathbf{c}_s \\ \mathbf{X}'_s \mathbf{w}_s - \mathbf{X}'_s \mathbf{c} \end{pmatrix} \right\|_V^2,$$

where

$$\mathbf{V} = \begin{pmatrix} \mathbf{U}_s & \mathbf{0} \\ \mathbf{0} & \alpha \mathbf{I} \end{pmatrix}$$

and $\alpha \geq 0$. We then minimize

$$\begin{aligned} & \|\mathbf{w}_s - \mathbf{A}_s \mathbf{c}_s\|_{\mathbf{U}_s}^2 + \alpha \|\mathbf{X}'_s \mathbf{w}_s - \mathbf{X}'_s \mathbf{c}\|_T^2 = \\ & D_s(\mathbf{w}_s) + \alpha \|\mathbf{X}'_s \mathbf{w}_s - \mathbf{X}'_s \mathbf{c}\|_T^2. \end{aligned}$$

A similar minimization problem is encountered with ridge regression. For $\alpha = 0$ the solution is provided by Horvitz-Thompson weights $\mathbf{w}_s = \mathbf{A}_s \mathbf{c}_s$. For $\alpha > 0$, we seek $\mathbf{w}_s(\alpha)$ minimizing $\|\mathbf{K}(\mathbf{w}_s - \mathbf{A}_s \mathbf{c}_s) - \mathbf{b}\|_V^2$, where $\mathbf{K} = (\mathbf{I}_n, \mathbf{X}_s)'$, $\mathbf{b} = (\mathbf{0}_{1 \times n}, (\mathbf{X}'_s \mathbf{c} - \mathbf{X}'_s \mathbf{A}_s \mathbf{c}_s)')'$ and $\mathbf{0}_{1 \times n} \in \mathbb{R}^n$ is a row vector of zeros. Ben-Israel and Greville (1980) yields

$$\mathbf{w}_s(\alpha) - \mathbf{A}_s \mathbf{c}_s = (\mathbf{K}' \mathbf{V} \mathbf{K})^{-1} \mathbf{K}' \mathbf{V} \mathbf{b}. \quad (10)$$

Thus by substituting the values of \mathbf{K} , \mathbf{V} , and \mathbf{b} we obtain

$$\begin{aligned} \mathbf{w}_s(\alpha) &= \mathbf{A}_s \mathbf{c}_s + \alpha (\mathbf{U}_s \\ &+ \alpha \mathbf{X}_s \mathbf{T} \mathbf{X}_s')^{-1} \mathbf{X}_s \mathbf{T} (\mathbf{X}'_s \mathbf{c} - \mathbf{X}'_s \mathbf{A}_s \mathbf{c}_s). \end{aligned} \quad (11)$$

It is easily shown that

$$\begin{aligned} & \alpha (\mathbf{U}_s + \alpha \mathbf{X}_s \mathbf{T} \mathbf{X}_s')^{-1} \mathbf{X}_s \mathbf{T} \\ &= \mathbf{U}_s^{-1} \mathbf{X}_s (\alpha^{-1} \mathbf{T}^{-1} + \mathbf{X}_s' \mathbf{U}_s^{-1} \mathbf{X}_s)^{-1}, \end{aligned}$$

hence

$$\begin{aligned} \mathbf{w}_s(\alpha) &= \mathbf{A}_s \mathbf{c}_s + \mathbf{U}_s^{-1} \mathbf{X}_s (\alpha^{-1} \mathbf{T}^{-1} \\ &+ \mathbf{X}_s' \mathbf{U}_s^{-1} \mathbf{X}_s)^{-1} (\mathbf{X}'_s \mathbf{c} - \mathbf{X}'_s \mathbf{A}_s \mathbf{c}_s). \end{aligned} \quad (12)$$

The estimator $Y'_s w_s(\alpha)$ thus becomes $\hat{Y}'c + (Y_s - \hat{Y}_s)'A_s c_s$, where $\hat{Y} = X\hat{\beta}_s(\alpha)$ and

$$\hat{\beta}_s(\alpha) = (X'_s U_s^{-1} X_s + \alpha^{-1} T^{-1})^{-1} X'_s U_s^{-1} Y_s.$$

The vector of regression coefficients, then, is the one obtained with ridge regression. Just as the calibration method, and the generalized regression method described by Särndal, Swensson and Wretman (1992), lead to the same estimators, a similar parallel can be drawn between mitigated calibration and ridge regression.

On the basis of equation (12), we can also use Ben-Israel and Greville (1980), and the fact that $F^\dagger = F'(FF')^\dagger$ with $F = T^{1/2} X'_s U_s^{-1/2}$, to show that

$$\lim_{\alpha \rightarrow \infty} w_s(\alpha) = w_{\text{cal}}.$$

This result was to be expected, since finding the vector $w_s(\alpha)$ which minimizes $D_s(w_s) + \alpha \|X'_s w_s - X'_s c\|_T^2$ when $\alpha \rightarrow \infty$ is equivalent to finding the weight vector which minimizes $D_s(w_s)$ among those which minimize $\|X'_s w_s - X'_s c\|_T^2$.

Rao and Singh (1997) defined tolerances for each of the p constraints of the calibration equation, and they established a relationship between these tolerances and the matrix αT .

For $\alpha \in [0, \infty[$ the function $w_s(\alpha)$ is represented by a curve in \mathbb{R}^n which links point $A_s c_s$ to point w_{cal} . If $p=1$, i.e. if X is a vector, this curve is a line segment. In fact, in this case the matrix $(\alpha^{-1} T^{-1} + X'_s U_s^{-1} X_s)^{-1}$ and the vector $X'_s c - X'_s A_s c_s$ are scalars, and the weights $w_s(\alpha)$ given by (12) are therefore equal to Horvitz-Thompson weights plus a multiple of vector $U_s^{-1} X_s$. And again for $p=1$, we have

$$\lim_{\alpha \rightarrow \infty} w_s(\alpha) = w_{\text{cal}} = A_s c_s + [(X'_s c - X'_s A_s c_s) / (X'_s U_s^{-1} X_s)] U_s^{-1} X_s$$

which leads to the estimator

$$Y'_s w_{\text{cal}} = Y'_s A_s c_s + [(Y'_s U_s^{-1} X_s) / (X'_s U_s^{-1} X_s)] (X'_s c - X'_s A_s c_s)$$

Taking $U = A^{-1} \text{diag}(X)$, we obtain the ratio estimator

$$Y'_s A_s c_s + [(Y'_s A_s \mathbf{1}_{n \times 1}) / (X'_s A_s \mathbf{1}_{n \times 1})] (X'_s c - X'_s A_s c_s),$$

where $\mathbf{1}_{a \times b} \in \mathbb{R}^{a \times b}$ is a matrix of ones.

Ben-Israel and Greville (1980, 111, exercise 15) showed that $D_s(w_s(\alpha))$ is an increasing monotonic function of α . Note however that for a unit $k \in s$, $|w_k(\alpha) - a_k c_k|$ is not necessarily a monotonic function of α . As α increases, the

weight vector $w_s(\alpha)$ moves away from the Horvitz-Thompson weight vector, but this does not necessarily apply to each coordinate of the vector.

In this article, mitigated calibration is used to restrict weights, i.e. when the size of the sample is relatively small. It can easily be shown, however, that for an asymptotic setup satisfying (2) and for which $\hat{\beta}_s(\alpha) - \beta(\alpha) \rightarrow 0$ in probability, with

$$\beta(\alpha) = (X'U^{-1}X + \alpha^{-1}T^{-1})^{-1}X'U^{-1}Y,$$

we have $Y'_s w_s(\alpha)$ is an asymptotically unbiased estimator whose asymptotic variance is

$$(Y - Y^*)' \text{diag}(c)(A\Pi A - \mathbf{1}_{N \times N}) \text{diag}(c)(Y - Y^*),$$

where $Y^* = X\beta(\alpha)$, Π is the matrix of inclusion probabilities of order 2, and $\text{diag}(c)$ is the diagonal matrix formed from vector c .

5. ESTIMATION METHODS WITH RESTRICTED WEIGHTS

In order to avoid obtaining weights having extreme values, we may wish to force the weight vector to be within a given region. This restricted region will be assumed to be convex and closed, and $A_s c_s$ will be assumed to be a point in this region. For example, if $w^{(L)} < A_s c_s < w^{(H)}$, we may wish to restrict the weights to region $R_w = \{w_s : w^{(L)} \leq w_s \leq w^{(H)}\}$. We will assume that

$$\lim_{n \rightarrow \infty} w^{(L)} - A_s c_s < 0 \quad \text{and} \quad \lim_{n \rightarrow \infty} w^{(H)} - A_s c_s > 0.$$

The approach described in section 3 consists in selecting a distance measure between calibrated weights and Horvitz-Thompson weights which will provide weights that satisfy the calibration equation and which lie in the restricted region, should such weights in fact exist. The approach dealt with in this section is to temperate the requirement that the calibration equation be satisfied when the vector of calibration weights w_{cal} is outside the restricted region. Various means to temperate this requirement lead to different weighting methods.

When w_{cal} lies outside the restricted region, we could for example look for those points on the curve $w_s(\alpha)$ parameterized by $\alpha \geq 0$ which are on the border of this region. One property of these points is that they solve the minimization problem described in section 4 for corresponding values of α , thus through matrix T , the importance of each calibration equation can be weighted. Using the example of the restricted region provided above, if

$$w_{\text{cal}} = \lim_{\alpha \rightarrow \infty} w_s(\alpha)$$

lies within this region, then $w_{\text{res } 1} = w_{\text{cal}}$ can be used as a restricted weight vector, otherwise $w_{\text{res } 1} = w_s(\alpha)$ with

$\alpha < \infty$ can be chosen such that $\mathbf{w}_s(\alpha)$ is on the boundary of the restricted region. If the asymptotic setup is such that conditions (2) are met with $\gamma < 3/2$ then for n sufficiently large, the probability that \mathbf{w}_{cal} will be within the restricted region is equal to one. In fact, we have $\mathbf{w}_{\text{cal}} - \mathbf{A}_s \mathbf{c}_s$ converging in probability to $\mathbf{0}$. The asymptotic properties of the estimator using the restricted weights, $\mathbf{w}_{\text{res } 1}$, are therefore identical to those of the calibration estimator. It is worth noting that since $|w_k - a_k c_k|$ is not necessarily a monotonic function of α , it is possible for $\mathbf{w}_s(\alpha)$ to be on the boundary of the restricted region for several values of α , even if the restricted region is convex. Finding all these values is not necessarily a simple matter, and a decision has to be taken as to which value to use.

Another option for restricting weights would be to use as a restricted region those weights \mathbf{w}_s which satisfy $D_s(\mathbf{w}_s) \leq l$ for a bound $l > 0$. Then $\mathbf{w}_{\text{res } 2} = \mathbf{w}_{\text{cal}}$ is taken as a restricted weight vector if \mathbf{w}_{cal} lies in the restricted region, otherwise we seek $\alpha > 0$ such that $D_s(\mathbf{w}_s(\alpha)) = l$. This value of α is unique and can be found through iteration. Next we calculate the weights $\mathbf{w}_{\text{res } 2} = \mathbf{w}_s(\alpha)$ which correspond to this value of α using equation (12). If the asymptotic setup is such that conditions (2) are met with $\gamma < 1$, and if l does not vary with n , then for n sufficiently large, the probability that \mathbf{w}_{cal} will be within the restricted region is equal to one. In fact, we have $D_s(\mathbf{w}_{\text{cal}})$ converging in probability to 0. The asymptotic properties of the estimator using restricted weights, $\mathbf{w}_{\text{res } 2}$, are then identical to those of the calibration estimator. Unfortunately, when estimating a total, we must expect to have $\gamma = 1$. In order to overcome this snag, we can use $l\sqrt{n}$ as a bound, instead of l . We can justify this bound on the basis that the length of the main diagonal of a hypercube of \mathbb{R}^n is equal to the diameter of the sphere which circumscribes this hypercube, whereas the diameter of the sphere inscribed in this same hypercube is smaller by a factor of \sqrt{n} . The fact remains that a statistician might be uncomfortable using an asymptotic setup where the bound increases with the size of the sample. Furthermore, with this approach, the weights of the observations cannot be limited individually. Only the distance between the restricted weight vector and the Horvitz-Thompson weight vector is controlled.

With the methods described above, we look for those points on curve $\mathbf{w}_s(\alpha)$ which lie on the boundary of the restricted region. The value of α for which $\mathbf{w}_s(\alpha)$ lies on the boundary of the restricted region must often be found iteratively. It would be simpler to replace the curve $\mathbf{w}_s(\alpha)$ by the line segment linking $\mathbf{A}_s \mathbf{c}_s$ to \mathbf{w}_{cal} . For the restricted region R_w , the restricted weight vector would be $\mathbf{w}_{\text{res } 3} = \mathbf{w}_{\text{cal}}$ if \mathbf{w}_{cal} is in the restricted region, otherwise $\mathbf{w}_{\text{res } 3}$ would be equal to the point at which the line segment crosses the boundary of restricted region, *i.e.*

$$\mathbf{w}_{\text{res } 3} = \mathbf{A}_s \mathbf{c}_s + \xi(\mathbf{w}_{\text{cal}} - \mathbf{A}_s \mathbf{c}_s),$$

where

$$\xi = \min_k \{ \max [(\mathbf{w}^{(L)} - \mathbf{A}_s \mathbf{c}_s) / (\mathbf{w}_{\text{cal}} - \mathbf{A}_s \mathbf{c}_s), (\mathbf{w}^{(H)} - \mathbf{A}_s \mathbf{c}_s) / (\mathbf{w}_{\text{cal}} - \mathbf{A}_s \mathbf{c}_s)] \},$$

vector division being elementwise, the maximum of the two vectors being elementwise, and \min_k providing the minimum element. We could also consider the weight vector of the restricted region, $\mathbf{w}_{\text{res } 4}$, which comes closest to \mathbf{w}_{cal} . Again for restricted region R_w , we would have

$$\mathbf{w}_{\text{res } 4} = \min [\max (\mathbf{w}_{\text{cal}}, \mathbf{w}^{(L)}), \mathbf{w}^{(H)}].$$

The asymptotic properties of estimators using restricted weights $\mathbf{w}_{\text{res } 3}$ or $\mathbf{w}_{\text{res } 4}$ are identical to those of the calibration estimator, as long as $\mathbf{w}_{\text{cal}} - \mathbf{A}_s \mathbf{c}_s$ converges in probability to $\mathbf{0}$, which is usually the case.

One interesting property of all the approaches discussed in this section is that, no matter what the restricted region, the existence of restricted weights is guaranteed. This is not always the case when using an approach based on distance measures. A simple example will now be introduced to allow comparisons between a few approaches.

We wish to estimate a total on the basis of a simple random sample of size 2 in a population of size 20. In other words, $\mathbf{c} = \mathbf{1}_{20 \times 1}$ and $\mathbf{a} = 10(\mathbf{1}_{20 \times 1})$. We use the auxiliary information vector $\mathbf{X} = (1, 2, 3, \dots, 20)'$, assume that the selected sample is $s = \{2, 12\}$ and choose U as a diagonal matrix with $u_{kk} = x_k = k$. A rectangular restricted region is provided using points $\mathbf{w}^{(L)} = (0, 0)'$ and $\mathbf{w}^{(H)} = (20, 13)'$. In other words, the weight of the first sample unit must be greater than 0 and less than 20, whereas the weight of the second sample unit must be greater than 0 and less than 13.

Under these conditions, the calibrated weights $\mathbf{w}_{\text{cal}} = (15, 5)'$ lie outside the restricted region. Since $p = 1$, weights $\mathbf{w}_s(\alpha)$ lie on the line segment which links $\mathbf{A}_s \mathbf{c}_s = (10, 10)'$ to \mathbf{w}_{cal} . We therefore have $\mathbf{w}_{\text{res } 1} = \mathbf{w}_{\text{res } 3}$, which means that the two methods give the same result. In this case, we have $\mathbf{w}_{\text{res } 1} = \mathbf{w}_{\text{res } 3} = (13, 13)'$. The method which consists in choosing that point in the restricted region which lies closest to the calibrated weights yields $\mathbf{w}_{\text{res } 4} = (15, 13)'$. On the other hand, if we look for $\mathbf{w}_{\text{res } 5}$, the restricted weights obtained while requiring that the calibration equation be satisfied and while using a distance measurement which assumes an infinite value outside the restricted region, then there is no solution. In fact, for any weight in the restricted region $\mathbf{X}'_s \mathbf{w}_s \leq 196$, whereas $\mathbf{X}' \mathbf{c} = 210$. If we had, say, $\mathbf{w}^{(H)} = (30, 13)'$, then using $D_s(\mathbf{w}_s)$ as a distance measurement within the restricted region we would have $\mathbf{w}_{\text{res } 5} = (27, 13)'$. These weights are fairly distant from $\mathbf{w}_{\text{cal}} = (15, 15)'$ and from $\mathbf{A}_s \mathbf{c}_s = (10, 10)'$. Such is the price to be paid for insisting on having weights which meet the calibration equation.

6. ESTIMATORS FOR DOMAINS WITH A SYNTHETIC COMPONENT

Restricted weights are used because of the properties of the calibration estimator for small sample sizes. For large sample sizes, we normally have $w_{\text{cal}} - A_s c_s$ converging in probability to $\mathbf{0}$, *i.e.* weights that are not problematic. A statistician faced with a problem of extreme weights must therefore in all likelihood also face another problem of small sample sizes, *i.e.* estimation for small domains. This section introduces an estimator whose asymptotic properties are those of the calibration estimator, but which uses restricted weights and takes advantage of a synthetic estimator.

Let $\tilde{Y} = X\tilde{\beta}_s$ denote a synthetic estimate for Y , we have

$$\begin{aligned}\tilde{Y}'w_s &= (X_s'\tilde{\beta}_s)'w_s \\ &= \tilde{\beta}_s'X_s'w_s \\ &\approx \tilde{\beta}_s'X'c \\ &= (X\tilde{\beta}_s)'c \\ &= \tilde{Y}'c\end{aligned}\quad (13)$$

with equality at the third step if the weights satisfy the calibration equation $X_s'w_s = X'c$. The weights w_{cal} given by (1) minimize $\|X_s'w_s - X'c\|_T^2$. We can therefore estimate $Y'c$ using

$$\hat{\tau} = (Y_s - \tilde{Y}_s)'w_{\text{res}} + \tilde{Y}'c. \quad (14)$$

There will be equality between this estimator and estimator $Y_s'w_{\text{cal}}$ once the sample is sufficiently large for the calibration equation to be satisfied and for w_{cal} to lie in the restricted region, *i.e.* once $w_{\text{res}} = w_{\text{cal}}$. The asymptotic properties of these two estimators are therefore identical under certain conditions discussed in the previous section. The advantage of using estimator $\hat{\tau}$ is that it provides a synthetic estimate when columns of Y_s and \tilde{Y}_s are zero.

7. OUTLIERS

Outliers could be dealt with in much the same way as extreme weights. The strategy is the following: we adopt a restricted region for $Y_s'w_{\text{cal}}$, we show that when n is sufficiently large $Y_s'w_{\text{cal}}$ lies within the restricted region, and we adopt a more "reasonable" estimator to replace $Y_s'w_{\text{cal}}$ in those cases where $Y_s'w_{\text{cal}}$ lies outside the restricted region. For a stratified sample, we would normally have one restricted region per stratum.

In section 2, it was shown that under certain conditions for the asymptotic setup, $w_{\text{cal}} - A_s c_s = O_p(n^{-3/2}N^\gamma)$. We thus have $Y_s'w_{\text{cal}} - Y_s'A_s c_s = O_p(n^{-1/2}N^\gamma)$, and if we assume that

$$Y_s'A_s c_s - Y'c = O_p(n^{-1/2}N^\gamma), \quad (15)$$

then $Y_s'w_{\text{cal}} - Y'c = O_p(n^{-1/2}N^\gamma)$. An expert (or a group of experts) could determine on the basis of information gathered independently of survey data that it would not be reasonable to have $Y_s'w_{\text{cal}}$ outside a certain region. If $Y'c$ lies within the restricted region (*i.e.* if the expert does not find it unreasonable to have an estimate of the parameter which would be equal to the true value, $Y'c$, of the parameter), if $\gamma = 0$, and if the restricted region does not vary with n or N (or if $\gamma = 1$, and the restricted region varies in proportion to N), then for sufficiently large n , the probability that $Y_s'w_{\text{cal}}$ will lie within the restricted region is equal to one. In those cases where $Y_s'w_{\text{cal}}$ lies outside the restricted region, we could use as an estimate the point in the restricted region that lies closest to $Y_s'w_{\text{cal}}$ or we could assume that the weight of the few observations that are deemed outliers is equal to one, and distribute their original weights (less the number of outliers) among the observations that are not outliers. The asymptotic properties of this modified estimator used to deal with outliers are then identical to those of the unmodified estimator.

In the case of a non-stratified sample, this method is relatively easy to apply. If however the sample is stratified, and if constraints are imposed on estimates for each stratum, then we have two additional problems. First, if the asymptotic setup is such that the number of strata increases in proportion to the size of the sample, then the assumption given in (15) does not hold, since the mean sample size per stratum remains constant as $n \rightarrow \infty$. We need to determine whether it is reasonable to adopt an asymptotic setup in which the number of strata is constant (or increases less rapidly than n). Such an asymptotic setup is less plausible if the number of observations per stratum is small. The second problem is linked to the difficulty for the expert to impose constraints on estimates for each of the strata. The greater the number of strata, the greater the risk that $Y'c$ will not lie in the restricted region defined by the expert. In fact, in the case of a stratified sample, it is preferable for the expert to use information that is independent of the survey data, in order to ensure strata homogeneity, prior to finalizing stratification. In other words, it is preferable to use the information available before the survey, in order to prevent outliers, rather than to correct them. If the information has been used in such a way that, before the survey, there is no reason to believe that there is any unrepresentative observation in any stratum, then there is no justification for assuming the opposite once the data have been collected.

8. CONCLUSION

If for large sample sizes the calibrated weights remain within a restricted region, then the asymptotic properties of the estimator with restricted weights are obviously identical to those of the calibration estimator. For a given asymptotic setup, we can usually expect to have $w_{\text{cal}} - A_s c_s$ converging in probability to 0, *i.e.* for sufficiently large sample sizes the calibrated weights w_{cal} will remain within the restricted region R_w if $A_s c_s$ lies within R_w . However, we have seen that for the estimate of a total, we do not necessarily have convergence to 0 for $D_s(w_{\text{cal}})$. We must therefore avoid having a restricted region defined by $\|w_s - A_s c_s\|_{V_s}^2 \leq l$ at least if we are estimating a total and not a mean.

We have provided necessary and sufficient conditions for the existence of weights restricted to intervals which satisfy the calibration equation. If such weights do not exist, the idea of satisfying the calibration equation exactly must be abandoned. The problem of calibration with restricted weights can be reformulated in such a way that a solution will always be possible. Some of the approaches described in this paper make it possible to obtain a solution without recourse to iterative methods. These are simple methods that are easy to interpret. The asymptotic properties of these estimators are usually identical to those of the calibration estimator without weight restrictions.

The problem of extreme weights is encountered for small sample sizes, thus the problem of estimating for small domains should be considered simultaneously. It is possible to take advantage of synthetic estimators while using an estimator with restricted weights having good asymptotic properties.

It is also possible to modify the calibration estimator, or any other asymptotically consistent estimator, so as to deal with outliers. The conditions under which this modified estimator will have asymptotic properties identical to those of the unmodified estimator are not easily verified, just as it is difficult to verify whether an outlier is in fact unrepresentative. However, such conditions make it possible to identify those factors which allow an estimator that is corrected for outliers to be statistically valid.

ACKNOWLEDGEMENT

The author wishes to thank an associate editor and a referee for constructive comments which have helped improve the paper.

APPENDIX A

We wish to verify that $\Omega(\phi) = l' (V\phi)_- - h' (V\phi)_+$ has a value of zero or less. First, it is easily shown that this is true for a vector ϕ , if and only if it is true for a vector $k\phi$ with arbitrary $k > 0$. Only the direction of ϕ matters. It is therefore sufficient to verify the condition for ϕ of norm

equal to one. For this proof, we will use the l_1 -norm of ϕ , $\|\phi\|_{l_1} = \sum_{i=1}^p |\phi_i|$. Vectors ϕ with $\|\phi\|_{l_1} = 1$ are located in hyperplanes whose intersections lie on points orthogonal to the unit vectors, *i.e.* points at least one of whose coordinates is zero. Function Ω varies linearly except at points ϕ orthogonal to one or more rows of V . Even when the domain of Ω is restricted to vectors ϕ with $\|\phi\|_{l_1} = 1$ that are orthogonal to $0 \leq j < (p-1)$ linearly independent rows of V , function Ω still varies linearly except at points orthogonal to other rows of V or orthogonal to unit vectors (which are likewise rows of V). The maximum of Ω for $\|\phi\|_{l_1} = 1$ is therefore reached at a point ϕ orthogonal to $(p-1)$ linearly independent rows of V . It is therefore sufficient to verify the condition for two vectors of opposite direction which are orthogonal to $(p-1)$ linearly independent rows of V , and this for each subset of $(p-1)$ linearly independent rows of V .

APPENDIX B

Let $\text{vec}(F)$ denote the vector obtained by piling successive columns of matrix $F \in \mathbb{R}^{a \times b}$ with the first column on top, and let the Kronecker product of two matrices F and G be defined as

$$F \otimes G = \begin{pmatrix} f_{11}G & \dots & f_{1n}G \\ \vdots & & \vdots \\ f_{m1}G & \dots & f_{mn}G \end{pmatrix}. \quad (\text{B1})$$

The result is derived from the corollary in section 3 with

$$M = \begin{pmatrix} I_{RC} \\ I_R \otimes \mathbf{1}_{1 \times C} \\ \mathbf{1}_{1 \times R} \otimes I_C \\ \mathbf{1}_{1 \times RC} \end{pmatrix}, \quad w = \text{vec}((\hat{N}_{ij}')'),$$

$$l = \begin{pmatrix} \text{vec}((N_{ij}^{(L)})') \\ N_{1.}^{(L)} \\ \vdots \\ N_{R.}^{(L)} \\ N_{.1}^{(L)} \\ \vdots \\ N_{.C}^{(L)} \\ N_{..}^{(L)} \end{pmatrix}, \quad h = \begin{pmatrix} \text{vec}((N_{ij}^{(H)})') \\ N_{1.}^{(H)} \\ \vdots \\ N_{R.}^{(H)} \\ N_{.1}^{(H)} \\ \vdots \\ N_{.C}^{(H)} \\ N_{..}^{(H)} \end{pmatrix}. \quad (\text{B2})$$

Only a finite set of conditions need be verified, first by noting that the columns of

$$V = \begin{pmatrix} -I_R \otimes \mathbf{1}_{C \times 1} & -\mathbf{1}_{R \times 1} \otimes I_C & -\mathbf{1}_{RC \times 1} \\ I_R & \mathbf{0}_{R \times C} & \mathbf{0}_{R \times 1} \\ \mathbf{0}_{C \times R} & I_C & \mathbf{0}_{C \times 1} \\ \mathbf{0}_{1 \times R} & \mathbf{0}_{1 \times C} & 1 \end{pmatrix} \quad (B3)$$

form a basis for $N(M')$. In other words, $M'V = \mathbf{0}$, the columns of V are linearly independent, and $N(M')$ is of dimension $R+C+1$. Note also that the last $R+C+1$ rows of V are the unit vectors. Finally, we verify the conditions of the corollary for all vectors $\lambda = V\phi$ and $\lambda = -V\phi$, where ϕ is orthogonal to $R+C$ linearly independent rows of V . This last step is described in greater detail in the following paragraph.

An arbitrary subset of $R+C$ linearly independent rows of V which includes the last row of V is denoted L , and the subset of all rows of V which are linear combinations of rows of L is denoted L^+ . If L^+ includes row $RC+i$ ($i = 1, \dots, R$) if and only if $i \notin S \subseteq \{1, 2, \dots, R\}$, and includes row $RC+R+j$ ($j = 1, \dots, C$) if and only if $j \notin T \subseteq \{1, 2, \dots, C\}$, then we set $\phi = (\phi'_S, -\phi'_T, 0)'$, where the i -th element of $\phi_S \in \mathbb{R}^R$ is equal to 1 if $i \in S$ and to 0 otherwise, and the j -th element of $\phi_T \in \mathbb{R}^C$ is equal to 1 if $j \in T$ and to 0 otherwise. Then

$$V\phi = ((-\phi'_S \otimes \mathbf{1}_{C \times 1} + \mathbf{1}_{R \times 1} \otimes \phi'_T)', \phi'_S, -\phi'_T, 0)',$$

therefore ϕ is orthogonal to all rows of L^+ , and all the more so ϕ is orthogonal to all rows of L . Likewise, vector $\phi^* = (\phi'_S, \phi'_T, -1)'$ is orthogonal to all rows of a subset of $R+C$ linearly independent rows of V which includes row $RC+i$ ($i = 1, \dots, R$) if and only if $i \notin S$, and includes row $RC+R+j$ ($j = 1, \dots, C$) if and only if $j \notin T$, but does not include the last row of V . The condition $-I'\lambda_- \leq h'\lambda_+$, with $\lambda = V\phi$ provides the fifth set of inequalities in (9). Likewise, by assuming λ equal to $-V\phi$, $V\phi^*$ and $-V\phi^*$ we obtain the last three sets of inequalities in (9).

REFERENCES

- BACHARACH, M. (1965). Estimating nonnegative matrices from marginal data. *International Economic Review*, 6, 294-310.
- BARDSLEY, P., and CHAMBERS, R.L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33, 290-299.
- BEN-ISRAEL, A., and GREVILLE, T.N.E. (1980). *Generalized Inverses: Theory and Applications*. Huntington, New York: Robert E. Krieger Publishing Company.
- BREWER, K.R.W. (1979). A class of robust sampling designs for large scale surveys. *Journal of the American Statistical Association*, 74, 911-915.
- DEVILLE, J.-C., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- DUCHESNE, P. (1999). Robust calibration estimators. *Survey Methodology*, 25, 43-56.
- FAN, K. (1956). On systems of linear inequalities. *Annals of Mathematics Studies*, (Eds. H. W. Kuhn, and A. W. Tucker), 38, 99-156.
- GRAYBILL, F.A. (1983). *Matrices with Applications in Statistics*, (Second Edition). Belmont, California: Wadsworth Publishing.
- ISAKI, C.T., and FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.
- RAO, J.N.K., and SINGH, A.C. (1997). A ridge-shrinkage method for range-restricted weight calibration in survey sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.
- SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- THÉBERGE, A. (1999). Extensions of calibration estimators in survey sampling. *Journal of the American Statistical Association*, 94, 635-644.

A Cautionary Note on Adjusting Weights for Nonresponse

WILLARD C. LOSINGER, LINDSEY P. GARBER, BRUCE A. WAGNER and GEORGE W. HILL¹

ABSTRACT

For surveys which involve more than one stage of data collection, one method recommended for adjusting weights for nonresponse (after the first stage of data collection) entails utilizing auxiliary variables (from previous stages of data collection) which are identified as predictors of nonresponse. In the final stage of data collection for the United States National Animal Health Monitoring System's Beef '97 Study, two variables were identified that clearly separated eligible producers by their propensity to respond. However, these variables were noticeably inferior to simple region by herd-size categories as predictors of responses that eligible producers gave for other questions in previous data-collection stages. Therefore, we decided to form weight-adjustment classes by region and herd size, even though other variables were greater predictors of response. When selecting auxiliary variables to adjust weights for nonresponse, we recommend that survey statisticians also evaluate the extent to which these auxiliary variables are related to data which nonrespondents would have provided. Using auxiliary variables which exhibit the greatest variation in response propensity may result in the greatest variation in weight-adjustment factors, but may bias population estimates for parameters unrelated to the chosen auxiliary variables.

KEY WORDS: Nonresponse bias; Response propensity; Logistic regression; National survey.

1. INTRODUCTION

In multistage surveys where some participants fail to respond during the final stage of data collection, one has considerable information about final-stage nonrespondents from previous stages of the survey. Rizzo, Kalton and Brick (1996) presented several methods for selecting auxiliary variables and adjusting weights for nonresponse when a large number of characteristics of the nonrespondents were known. These methods concentrated on identifying and using characteristics that discriminated between respondents and eligible nonrespondents. However, by adjusting weights based on specific variables which demonstrate the greatest difference in response rates, one may potentially introduce bias in the survey estimates if these variables are unrelated to responses that would have been given by nonrespondents during the final stage of data collection. Therefore, one should also utilize data from the previous stages of data collection to determine whether the chosen auxiliary variables are linked to other characteristics of those eligible to participate in the survey.

The Beef '97 Study (of the National Animal Health Monitoring System (NAHMS) of the United States Department of Agriculture (USDA)) took place in 23 states and involved three stages of data collection. In the first stage (December 30, 1996 through February 3, 1997), enumerators from the USDA: National Agricultural Statistics Service collected data on general management practices from 2,713 agricultural operations with one or more beef cows. First-stage respondents who had five or more beef cows on January 1, 1997 were eligible to continue in the

second stage of data collection (from March 3 through May 23, 1997), provided they had at least one beef cow and remained in business at the time of the second stage of data collection. A total of 1,190 producers participated in the second stage of data collection, which involved an on-farm visit by a veterinary medical officer or animal health technician and concentrated on the health management of the beef cattle.

All operations that participated in the second stage of data collection were eligible to participate in the third and final stage of data collection (August 1, 1997 through January 31, 1998). A total of 952 (80.0%) eligible operations responded in the final stage. From the first two stages of data collection, a considerable amount of information was available on the 238 nonrespondents for the final stage of data collection. The purpose of this note is to describe the methods that were evaluated for adjusting the sample weights for nonresponse in the final stage of data collection for the NAHMS Beef '97 Study.

In addition to region and herd-size (based on the number of beef cows) categories, 45 variables based on data collected during the first two stages of interviews were evaluated for their impact on final-stage response rates. A stepwise variable selection procedure, with region and herd size forced into a logistic regression model and a significance level of 0.05 for other variables to enter and remain in the model, was used (Table 1). The logistic regression analysis demonstrated that there were some differences in final-stage response by region, but that differences in response by herd size were not significant. Increased nonresponse was associated with having only one breeding

¹ W.C. Losinger, L.P. Garber, B.A. Wagner and G.W. Hill, United States Department of Agriculture, Animal and Plant Health Inspection Service, Veterinary Services, Center for Epidemiology and Animal Health, 555 South Howes Street, Suite 200, Fort Collins, Colorado 80521 U.S.A.

season and not consulting a veterinarian to treat or diagnose disease during 1996. The potential use of the logistic-regression variables as auxiliary variables in creating cells to adjust weights for final-stage nonresponse was examined. Four categorization schemes for nonresponse weight adjustment were proposed:

1. The traditional region by herd size scheme with 15 cells.
2. Region by herd size except in the West, which was subdivided by the number of breeding seasons, for a grand total of 14 cells.
3. Subdividing the cells of option 2 (by either of the auxiliary variables) if the difference in response rate (between the two new subdivisions) was at least ten percent and at least 20 respondents remained in each cell. Two subdivisions occurred, which yielded a total of 16 cells.
4. Continuing the subdivision of categories, based on the greatest difference in response rate, until a minimum number of respondents (no fewer than 20) remained in each cell. This yielded a total of 24 cells.

Table 1

Results of Stepwise Logistic Regression to Identify Variables Associated With Nonresponse to the Final Stage of Data Collection for the National Animal Health Monitoring System's Beef '97 Study. Based on 1,190 Eligible Operations and 238 Nonrespondents

Variable/ Response	Parameter Estimate	P
Intercept	0.369	0.181
Region		
Northcentral	0.851	0.000
Southcentral	0.822	0.000
Central	2.062	0.000
Southest	1.164	0.000
West	1.000	
Number of beef cows		
1 - 49	0.299	0.106
50 - 99	0.146	0.151
100 +	1.000	
Number of breeding Seasons		
1	-.370	0.039
>1 or no set season	1.000	
A veterinarian was consulted to treat or diagnose disease in 1996		
Yes	0.441	0.005
No	1.000	

Adjustment factors for weights of final-stage respondents were computed by dividing the sum of second-stage weights for eligible operations by the sum of second-stage weights for final-stage respondents within each cell.

Since the establishment of cells for schemes 2 through 4 was based on variables which demonstrated the greatest differences in response rates, differences in adjustment factors increased for particular subcategories from scheme 1 to scheme 4. For example, for the first scheme, adjustment factors for the Western region were 1.897, 1.504 and 1.579 for the small, medium and large herd size categories respectively. For the second scheme, adjustment factors in the Western region were 1.334 for operations that did not have one defined breeding season, and 1.875 for operations that did have one defined breeding season. For the third scheme, operations in the West that had one defined breeding season were split into two cells based on whether they had used a veterinarian to diagnose or treat disease during 1996: operations that had indicated "yes" received a weight adjustment of 1.548, while operations that had indicated "no" received a weight adjustment of 2.326.

To investigate how well the proposed auxiliary variables might have related to overall management strategies, we selected additional variables from the first two stages of data collection, and, within each region, examined differences in these variables by herd size category, number of breeding seasons, and whether a veterinarian had been consulted to diagnose or treat disease during 1996. Table 2 presents some representative results for the Western region. Some herd-size differences existed in the percent of operations that had one set breeding season and the percent of operations that had consulted a veterinarian during 1996. However, the percent of operations that had consulted a veterinarian was practically identical for operations that had one set breeding season versus operations that did not have one set breeding season, and vice versa. In addition, the percent of operations that vaccinated heifers for brucellosis and the percent of operations that implanted calves with a growth promotant exhibited a wider range by herd size category than by the other two proposed auxiliary variables. Moreover, mean weaning age and mean calf death loss varied more by herd size than by either number of breeding seasons or by whether a veterinarian was consulted. Similar patterns were noticed for other regions.

Although herd size was not a statistically significant predictor of participation in the final stage of data collection for the NAHMS Beef '97 Study (table 1), herd size was found to be more highly related to a number of questionnaire variables than either of the additional proposed auxiliary variables which derived from the logistic regression analysis. Therefore, we utilized the traditional region by herd size category scheme to perform the nonresponse weight adjustment for the final stage of data collection for the NAHMS Beef '97 Study.

Table 2

For 261 Western-Region Operations Eligible to Participate in the Third and Final Phase of Data Collection for the United States National Animal Monitoring System's 1997 Beef '97 Study (August 1 through January 31, 1998), Responses to Selected Variables From the First two Phases of Data Collection by Auxiliary Variables Examined for Weight Adjustment for the Final Stage of Data Collection

Auxiliary variables proposed for weight adjustment for third-stage nonresponse	Variables selected from the first two stages of data collection					
	1	2	3	4	5	6
	Percent			Mean		
Number of beef cows						
1 - 49	69.2	50.8	63.1	15.4	215	6.3
50 - 99	69.2	59.6	80.8	26.9	232	3.9
100+	88.2	70.1	85.4	52.8	223	4.1
Number of breeding seasons						
1	—	62.3	69.8	17.0	223	5.1
>1 or no set season	—	63.5	81.3	43.8	223	4.5
A veterinarian was consulted to treat or diagnose disease in 1996						
Yes	79.2	—	69.8	28.1	222	4.5
No	80.0	—	84.2	44.2	223	4.6

Variables selected from the first two phases of data collection:

- 1 = Operations with one set breeding season
- 2 = Operations that consulted a veterinarian to treat or diagnose disease in 1996
- 3 = Operations that vaccinate any heifers for brucellosis
- 4 = Operations that implanted any calves with a growth promotant prior to or at weaning during 1996
- 5 = Average age (in days) of calves at weaning
- 6 = Percent of calves that died in 1996

Researchers using survey data depend on sample weights to produce population parameter estimates that are approximately unbiased. In the final stage of data collection for the NAHMS Beef '97 Study, a logistic regression analysis identified two variables that were superior to herd size as

predictors of nonresponse in the final stage of data collection. However, these variables were generally inferior to herd size in differentiating how producers responded to a number of key questions related to operation management. Using these two variables to establish categories for weight adjustment for nonresponse could have reduced bias in estimates of parameters (from the third stage of data collection) with which they were correlated. However, estimates of parameters not correlated with these variables could have been distorted. Therefore, we chose the traditional approach of performing the nonresponse weight adjustment by region and herd size categories.

Identifying variables that are good predictors of panel nonresponse is a good practice in any multistage survey. Prior to using these variables to adjust weights for unit nonresponse, we recommend that survey statisticians first follow some procedures to determine the extent to which these variables are linked to other characteristics of those eligible to complete the survey. Adjusting the weights based solely on variables that prove to be good predictors of panel nonresponse could potentially result in warped population estimates if these variables are not also good predictors of data that nonrespondents would have provided on the survey instrument.

ACKNOWLEDGEMENTS

The authors are grateful to the National Agricultural Statistics Service who initially selected the sample and the National Agricultural Statistics Service enumerators who made the first on-farm contact; the federal and state veterinarians and animal health technicians who made the subsequent on-farm visits; and all of the eligible beef producers, both respondents and nonrespondents.

REFERENCE

- RIZZO, L., KALTON, G., and BRICK, J. M. (1996). A comparison of some weighting adjustment methods for panel nonresponse. *Survey Methodology*, 22, 43-53.

Local Unconditional Best Linear Unbiased Estimators: Applications to Survey Sampling

JULIET POPPER SHAFFER¹

ABSTRACT

Survey statisticians frequently use superpopulation linear regression models. The Gauss-Markov theorem, assuming fixed regressors or conditioning on observed values of regressors, asserts that the standard estimators of regression coefficients are best linear unbiased. Shaffer (1991) showed that the Gauss-Markov theorem doesn't apply when the regressors are random if some aspects of the population distribution of the regressors are known, and introduced an alternative estimator with better properties than the standard estimator under some conditions. This paper derives some generalizations, and notes an optimality property (locally best linear unbiasedness) of the generalized alternative estimator. Implications for estimation in surveys are noted.

KEY WORDS: Regression analysis; Gauss-Markov theorem; Survey sampling; Unbiased estimation; Optimality; Best linear unbiased estimation.

1. INTRODUCTION

In the standard linear regression model for a sample of observations,

$$Y = X\beta + \epsilon, \quad (1)$$

the matrix of regressors, X , is assumed to be a known, fixed matrix. Shaffer (1991) showed that when X is assumed to be random, the Gauss-Markov theorem does not hold in general, and described an alternative estimator that is more accurate when β is close to zero. Shaffer gave two applications of her results, to estimates of β and associated population quantities in multivariate normal superpopulation models and to ratio estimation of population means and totals.

In the present paper, three generalizations of these results are derived.

- (a) The results are generalized from a model in which the sample covariance matrix of the errors ϵ is $\sigma^2 I$, where I is the $n \times n$ identity matrix, to the case in which the covariance matrix \sum of ϵ is $\sigma^2 B$, where B is a known, fixed positive-definite matrix, and to some situations in which B is random (since it is the covariance matrix of a randomly-selected sample of regressor values).
- (b) A generalized estimator is derived that performs well when the coefficient vector β is close to any pre-specified coefficient vector β_0 .
- (c) A condition is given for design-unbiasedness of estimators of population means and totals based on the generalized estimator of β .

Some results under the general model (1) will be given first. Then, modifications that apply to the sample survey situation will be discussed.

Under Model (1) with $\sum = \sigma^2 I$, the Gauss-Markov theorem asserts that the sample estimator

$$\hat{\beta} = (X'X)^{-1}X'Y, \quad (2)$$

is a best linear unbiased estimator (BLUE) if X is regarded as a fixed matrix. If the rows of X are treated as realizations of random vectors $x_i, i = 1, \dots, n$, the Gauss-Markov theorem can be interpreted as an assertion that the estimator in (2) has minimum variance in the class of estimators linear in Y and conditionally unbiased, given these realized values of X . However, the use of the term "unbiased" without qualification generally means unconditional unbiasedness. If the requirement of unbiasedness is interpreted to mean unbiased unconditionally, *i.e.*, on the average over random vectors with values in X , Shaffer (1991) showed that the Gauss-Markov theorem doesn't apply when $E(X'X)$ is known. In that case, the conditionally biased estimator

$$\hat{\beta}^* = [E(X'X)]^{-1}(X'Y) \quad (3)$$

is unconditionally unbiased and has smaller variance than $\hat{\beta}$ when β is small. In fact, when $E(X'X)$ is known, no BLUE exists.

Comparison of the variances of (2) and (3) under various modeling assumptions, aside from the implications for estimating the coefficients themselves, gives insight into the conditions under which various estimators of other parameters of the populations have desirable properties, both model-based and design-based.

¹ Juliet Popper Shaffer, University of California, Department of Statistics, 367 Evans Hall, #3860, Berkeley, CA 94720-3860, U.S.A.
E-mail: shaffer@berkeley.edu.

2. GENERALIZATION OF THE COVARIANCE MATRIX OF ϵ

If the covariance matrix of ϵ is of the form $\sigma^2 \mathbf{B}$, where \mathbf{B} is a known, fixed positive-definite matrix, the Gauss-Markov theorem applies to the generalized estimator

$$\hat{\beta} = [\mathbf{X}'\mathbf{B}^{-1}\mathbf{X}]^{-1} \mathbf{X}'\mathbf{B}^{-1} \mathbf{Y}. \quad (4)$$

The proofs in Shaffer (1991) generalize directly to show that, if $\mathbf{E}(\mathbf{X}'\mathbf{B}^{-1}\mathbf{X})$ is known, the estimator

$$\hat{\beta}^* = [\mathbf{E}(\mathbf{X}'\mathbf{B}^{-1}\mathbf{X})]^{-1} \mathbf{X}'\mathbf{B}^{-1} \mathbf{Y} \quad (5)$$

has smaller variance than (4) when β is sufficiently close to zero. The (unconditional) variances of (4) and (5) are

$$\sum_{\hat{\beta}} = \mathbf{E}[(\mathbf{X}'\mathbf{B}^{-1}\mathbf{X})^{-1}] \sigma^2 \quad (6)$$

and

$$\begin{aligned} \sum_{\hat{\beta}^*} &= [\mathbf{E}(\mathbf{X}'\mathbf{B}^{-1}\mathbf{X})]^{-1} \sigma^2 \\ &+ \text{Var.}\{[\mathbf{E}(\mathbf{X}'\mathbf{B}^{-1}\mathbf{X})]^{-1} (\mathbf{X}'\mathbf{B}^{-1}\mathbf{X})\beta\}. \end{aligned} \quad (7)$$

When $\beta = 0$, Shaffer shows that (7) is smaller than (6), and therefore, assuming continuity of (7) as a function of β , it is smaller than (6) when β is in a neighborhood of zero.

The results will now be applied in the sample survey context. Let \mathbf{X}_N refer to the $N \times p$ matrix, and \mathbf{Y}_N to the $N \times 1$ vector, in a finite population. If the N population elements are considered to be a sample from an infinite hypothetical population of potential elements satisfying (1), and if a sample of size n of the finite population is taken, the proofs in Shaffer (1991) generalize directly to show that

$$\hat{\beta}_N^* = [\mathbf{E}(\mathbf{X}_N' \mathbf{B}_N^{-1} \mathbf{X}_N)]^{-1} \mathbf{X}_N' \mathbf{B}_N^{-1} \mathbf{Y}_N \quad (8)$$

and

$$\hat{\beta}_n^* = [\mathbf{E}(\mathbf{X}_n' \mathbf{B}_n^{-1} \mathbf{X}_n)]^{-1} \mathbf{X}_n' \mathbf{B}_n^{-1} \mathbf{Y}_n \quad (9)$$

have variances smaller than those of their corresponding conditional versions $\hat{\beta}_N$ and $\hat{\beta}_n$ respectively, if β is close to zero, where the expectation in (8) is over the infinite population of hypothetical elements, and the expectation in (9) is over either the same infinite population or over the finite population of N elements satisfying (1). In order to apply these results, the expectations in (8) and (9) have to be known.

If \mathbf{X}_N is to be regarded as fixed, the population model can be written as

$$\mathbf{Y}_N = \mathbf{X}_N \beta + \epsilon_N, \quad (10)$$

where ϵ_N is a vector of randomly distributed error terms as in (1). Under Model (10), $\hat{\beta}_N$ and $\hat{\beta}_N^*$ are identical, but $\hat{\beta}_n$ is still distinct from $\hat{\beta}_n^*$. Under Model (10), for a random sample of size n , if

$$\mathbf{E}[(\mathbf{X}_n' \mathbf{B}_n^{-1} \mathbf{X}_n)/n] = (\mathbf{X}_N' \mathbf{B}_N^{-1} \mathbf{X}_N)/N, \quad (11)$$

the alternative estimator can be written in the form

$$\hat{\beta}_n^* = [(n/N)(\mathbf{X}_N' \mathbf{B}_N^{-1} \mathbf{X}_N)]^{-1} \mathbf{X}_n' \mathbf{B}_n^{-1} \mathbf{Y}_n. \quad (12)$$

In model (10), Equations (11) and (12) will apply if \mathbf{B}_N is diagonal and the sampling plan is self-weighting, and under some other conditions and sampling plans, e.g., if \mathbf{B}_N is block (cluster) diagonal and complete clusters are sampled. If \mathbf{B}_N is diagonal, \mathbf{B}_n is not necessarily fixed. For example, suppose a population consists of both men and women, and the variances of the two sexes on the characteristic of interest are known and are different. In that case, if a self-weighting sample is taken, and Model (10) is assumed to hold in both subpopulations, \mathbf{B}_n will be diagonal, with entries that are a function of the proportions of the two genders in the sample.

3. LOCALLY BEST LINEAR UNBIASED ESTIMATION

Under the model (1), the estimator (5) is the locally best linear unbiased estimator (LBLUE) when $\beta = 0$; i.e., the estimator, linear in \mathbf{Y} and unbiased for β with smallest variance in a neighborhood of $\beta = 0$. Furthermore, the generalized linear estimator

$$\hat{\beta}_{(\beta_0)}^* = \beta_0 + [\mathbf{E}(\mathbf{X}'\mathbf{B}^{-1}\mathbf{X})]^{-1} [\mathbf{X}'\mathbf{B}^{-1}(\mathbf{Y} - \mathbf{X}\beta_0)], \quad (13)$$

allowing for the addition of a constant, is the LBLUE at $\beta = \beta_0$, for an arbitrary vector β_0 . The proof of these results is given in Appendix A. This generalized estimator (13) could be useful in a survey sampling situation in which it was reasonably sure that β would be close to some specified value. The variance of (13) is easily shown to equal (7) with $(\beta - \beta_0)$ substituted for β . (See Appendix A.) Under Model (10) estimators (8), (9), and (12) generalize to

$$\hat{\beta}_{(\beta_0, N)}^* = \beta_0 + [\mathbf{E}(\mathbf{X}_N' \mathbf{B}_N^{-1} \mathbf{X}_N)]^{-1} [\mathbf{X}_N' \mathbf{B}_N^{-1} (\mathbf{Y}_N - \mathbf{X}_N \beta_0)], \quad (14)$$

$$\hat{\beta}_{(\beta_0, n)}^* = \beta_0 + [\mathbf{E}(\mathbf{X}_n' \mathbf{B}_n^{-1} \mathbf{X}_n)]^{-1} [\mathbf{X}_n' \mathbf{B}_n^{-1} (\mathbf{Y}_n - \mathbf{X}_n \beta_0)], \quad (15)$$

and

$$\hat{\beta}_{(\beta_0, n)}^* = \beta_0 + [(n/N)\mathbf{X}_N' \mathbf{B}_N^{-1} \mathbf{X}_N]^{-1} [\mathbf{X}_n' \mathbf{B}_n^{-1} (\mathbf{Y}_n - \mathbf{X}_n \beta_0)], \quad (16)$$

respectively.

4. CONDITIONS FOR DESIGN UNBIASEDNESS

Assume the Model (10) holds, and that the unconditionally unbiased estimator can be expressed in the form (16). Suppose there exists a $p \times 1$ -vector \mathbf{g} such that $\mathbf{B}_N^{-1} \mathbf{X}_N \mathbf{g} = \mathbf{1}_N$ and, for every sample of size n , $\mathbf{B}_n^{-1} \mathbf{X}_n \mathbf{g} = \mathbf{1}_n$, where $\mathbf{1}_N$ and $\mathbf{1}_n$ are vectors of ones of length N and n , respectively. Then, given a simple random sample,

(a) the estimator

$$\hat{\bar{Y}}_{(\beta_{0,n})} = \bar{X}_N' \hat{\beta}_{(\beta_{0,n})}^* \quad (17)$$

is a design-unbiased estimator of \bar{Y}_N , where $\bar{X}_N' = (1/N) \mathbf{1}_N' \mathbf{X}_N$, and

(b) $\hat{\bar{Y}}_{(\beta_{0,n})}$ is a generalized difference estimator of \bar{Y}_N .

The proof is given in Appendix B.

Note that a vector \mathbf{g} satisfying the conditions of this theorem exists if the model includes an intercept (*i.e.*, \mathbf{X}_N includes a column of ones) or if \mathbf{B}_N is diagonal and the variance is proportional to the values of one of the regressors. Many applications of regression modeling to sample survey estimation are based on models that incorporate these assumptions. Särndal, Swensson and Wretman (1991, p. 231 and 232) discuss these and more general models, and Chapter, 6, section 4 of that reference has examples of commonly applied models incorporating these assumptions. Chapter 6 as a whole discusses both the general difference estimator of $N\bar{Y}_N$ and the analogous general regression estimator based on $\hat{\beta}_n$. The material in that Chapter also suggests generalizations of these results to more complex estimators and sampling plans.

5. DISCUSSION

To apply the results to estimates of properties of a finite population, it will be assumed that the matrix \mathbf{B} is diagonal or has the special block-diagonal form and associated sampling plan discussed above. From the results in section 3, it follows that the estimator (17) of \bar{Y}_N has smaller variance than the estimator

$$\hat{\bar{Y}}_{(\hat{\beta}_n)} = \bar{X}_N' \hat{\beta}_n \quad (18)$$

when β is close to β_0 . Note that (18) can be written

$$\hat{\bar{Y}}_{\hat{\beta}_n} = \frac{1}{N} \left[\sum_{i \in s} \mathbf{X}_i' \hat{\beta}_n + \sum_{i \notin s} \mathbf{X}_i' \hat{\beta}_n \right], \quad (19)$$

and \mathbf{X}_i' is the i -th row of \mathbf{X} , and S is the set of elements in the sample. Royall (1970) showed that the best linear

model-unbiased estimator of \bar{Y}_N (unbiased conditionally on the obtained sample) is

$$\frac{1}{N} \left[\sum_{i \in s} \mathbf{Y}_i + \sum_{i \notin s} \mathbf{X}_i' \hat{\beta}_n \right]. \quad (20)$$

In some important cases, the first term in (20) is equal to the first term in (19), in which case (20) and (19) are identical. This will be true, for example, if $\mathbf{B} = \sigma^2 \mathbf{I}$ and the model (10) contains an intercept, or if $p = 1$ and \mathbf{B} is diagonal with diagonal entries proportional to the values of the single regressor. In such cases, (20) and (19) are identical, and the design-unbiased and unconditionally-model-unbiased estimator (17) has a smaller expected squared discrepancy from \bar{Y}_N than the best linear conditionally-model-unbiased estimator (20) when β is close to β_0 . Furthermore, if the sampling fraction is negligible, (17) has smaller expected squared discrepancy than (20) when β is close to β_0 , even without the requirement that the first terms of (20) and (19) be equal.

If $\hat{\beta}$ is replaced by $\hat{\beta}^*$ in (20), the resulting estimator is no longer unconditionally unbiased. It can be shown, however, using concepts of dependence (Lehmann, 1966) that under the conditions on \mathbf{B} noted at the beginning of this section, the resulting estimator will have smaller expected squared discrepancy from \bar{Y}_N than (20) and (19) even without the further restrictions noted in the previous paragraph.

6. CONCLUSION

Since the conditions under which the estimator (5) of β is more efficient than the estimator (4) are very restrictive, and the estimators of population characteristics based on (5) can be derived in other ways, the results given here may be of more theoretical than practical interest. The results do give additional insight into some situations in which simple estimators like the sample mean and the generalized difference estimator are more efficient in estimating the population mean than are ratio estimators, poststratified estimators, regression estimators and other complex estimators. The equations (6) and (7) for comparative variances of (4) and (5) provide an alternative method of comparing respective variances under different regression models and different values of β . Many of these results hold under very simple sampling plans, but it should be possible to generalize them to more complex, unequal probability sampling plans.

ACKNOWLEDGEMENTS

The author is grateful to the late Erik N. Torgersen, who suggested the generalized estimator with optimal properties, to Phillip S. Kott, whose suggestion led to the derivation of the design unbiasedness condition, and to anonymous referees for many valuable comments.

APPENDIX A

Proof that $\hat{\beta}_{\beta_0}^*$ is LBLUE at β_0

Assume model (1), with $\text{Var}(Y|X) = \sigma^2 B$. (The general proof given here applies directly to the model (10) as well.) Consider the sample estimator

$$\hat{\beta}_{(\beta_0)}^* = \beta_0 + [E(X'B^{-1}X)]^{-1}X'B^{-1}(Y - X\beta_0).$$

Let $\tau = \beta - \beta_0$ and $Z = Y - X\beta_0$. Then $E(Z|X) = X\tau$, $\text{Var}(Z|X) = \sigma^2 B$, and $\hat{\tau}^* = [E(X'B^{-1}X)]^{-1}X'B^{-1}Z = \hat{\beta}_{(\beta_0)}^* - \beta_0$.

Thus, the properties of $\hat{\beta}_{(\beta_0)}^*$ at $\beta = \beta_0$ are the same as those of $\hat{\beta}^* = \hat{\beta}_{(0)}^*$ at $\beta = 0$, so without loss of generality it will be shown that $\hat{\beta}_{(0)}^*$ is LBLUE at $\beta = 0$. Also without loss of generality, it will be assumed that $B = I$.

Let $C'(X)Y$ be an arbitrary unconditionally-unbiased estimator of β , where $C(X)$ is a matrix of functions of X , of the same dimensions as X . The requirement of unconditional unbiasedness necessitates the restriction $E[C'(X)X] = I$ (Shaffer 1991). Conditioning first on X and then using the expression for unconditional variance, the variance of $C'(X)Y$ is $E[C'(X)C(X)]\sigma^2 + \text{Var}(C'(X)X\beta)$. Since we are considering variance at $\beta = 0$, only the first term is nonzero. Letting $C'(X) = [E(X'X)]^{-1}X'$, the variance of $\hat{\beta}^*$ is $[E(X'X)]^{-1}\sigma^2$.

Let $\tilde{\beta}$ be an arbitrary unconditionally-unbiased estimator of the form $C'(X)Y$. Then $\text{Var}(\tilde{\beta}) = \text{Var}(\hat{\beta}^*) + \text{Var}(\tilde{\beta} - \hat{\beta}^*) + 2\text{Cov}(\hat{\beta}^*, \tilde{\beta} - \hat{\beta}^*)$, so $\text{Var}(\hat{\beta}^*) \leq \text{Var}(\tilde{\beta})$ if $\text{Cov}(\hat{\beta}^*, \tilde{\beta} - \hat{\beta}^*) \geq 0$, or if $\text{Cov}(\hat{\beta}^*, \tilde{\beta}) \geq \text{Var}(\hat{\beta}^*)$. An easy calculation, using the restriction $E[C'(X)X] = I$, shows that $\text{Cov}(\hat{\beta}^*, \tilde{\beta}) = \text{Var}(\hat{\beta}^*)$, which proves that $\hat{\beta}_{(\beta_0)}^*$ is LBLUE at β_0 .

APPENDIX B

Proof of the Result in Section 4

$$\bar{X}_N' \hat{\beta}_{(\beta_0)}^* = \bar{X}_N' \beta_0 + (1/N) \mathbf{1}_N' X_N [n(1/N)(X_N' B_N^{-1} X_N)^{-1}]$$

$$X_n' B_n^{-1} (Y_n - X_n \beta_0)$$

$$= \bar{X}_N' \beta_0 + (1/n) g' X_n' B_N^{-1} X_N (X_N' B_N^{-1} X_N)^{-1}$$

$$X_n' B_n^{-1} (Y_n - X_n \beta_0)$$

$$= \bar{X}_N' \beta_0 + (1/n) \mathbf{1}_n' (Y_n - X_n \beta_0)$$

$$= \bar{X}_N' \beta_0 + \bar{Y}_n - \bar{X}_n \beta_0. \quad (\text{B.1})$$

where B_N and B_n are the appropriate population and sample matrices, respectively. The final expression in (B.1) is the generalized difference estimator based on a value β_0 chosen independently of the sample. This proves part (b) of the result; since the difference estimator is unbiased for \bar{Y} in a self-weighting sample, the result in (a) follows.

REFERENCES

- LEHMANN, E.L. (1966). Some concepts of dependence. *Annals of Mathematical Statistics*, 37, 1137-1153.
- ROYALL, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.
- SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1991). *Model Assisted Survey Sampling*. New York: Springer.
- SHAFFER, J.P. (1991). The Gauss-Markov theorem and random regressors. *The American Statistician*, 45, 269-273.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board

Contents Volume 15, Number 4, 1999

Bayesian Estimation of the Number of Unseen Studies in a Meta-Analysis <i>Lynn E. Eberly and George Casella</i>	477
Toward a Social Psychological Programme for Improving Focus Group Methods of Developing Questionnaires <i>Katherine Bischooping and Jennifer Dykema</i>	495
Statistical Methods for Developing Ratio Edit Tolerances for Economic Data <i>Katherine Jenny Thompson and Richard S. Sigman</i>	517
A Conditional Analysis of Some Small Area Estimators in Two Stage Sampling <i>Piero D. Falorsi and Aldo Russo</i>	537
Internal Migration: What Data are Available in Europe? <i>Philip Rees and Marek Kupiszewski</i>	551
A Bibliography on Statistical Consulting and Training <i>Hardeo Sahai and Anwer Khurshid</i>	587
Editorial Collaborators	631
Index to Volume 15, 1999	663

All inquires about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Department, Statistics Sweden, Box 24 300, S - 104 51 Stockholm, Sweden.

CONTENTS

TABLE DES MATIÈRES

Volume 27, No. 4, December/décembre 1999

Jerald F. LAWLESS Statistical science: concepts, opportunities and challenges	671
Byron SCHMULAND Dirichlet forms: some infinite-dimensional examples	683
Joseph G. IBRAHIM, Ming-Hui CHEN and Steven N. MacEachern Bayesian variable selection for proportional hazards models	701
Yodit SEIFU, Thomas A. SEVERINI and Martin A. TANNER Semiparametric Bayesian inference for regression models	719
Konstantinos FOKIANOS, Amy PENG and Jing QIN A generalized-moments specification test for the logistic link	735
Zhide FANG and Douglas P. WIENS Robust extrapolation designs and weights for biased regression models with heteroscedastic errors	751
Michael P. JONES Nonrobustness of the information test in detecting heterogeneity	771
Douglas P. WIENS and Julie ZHOU Minimax designs for approximately linear models with AR (1) errors	781
Luc D. ADJENGUE and Marc MOORE Deux méthodes d'estimation pour les paramètres de processus moyenne mobile spatiaux	795
Benoît R. MÂSSE and Young K. TRUONG Conditional logspline density estimation	819
Satish IYENGAR, Paul KVAM and Harshinder SINGH Fisher information in weighted distributions	833
E.G. ENNS, P.F. EHLERS and T. MISI A cluster problem as defined by nearest neighbours	843
Mohammadine BELBACHIR Lois limites pour les statistiques d'ordre dans le cas non identiquement distribué	853
Bradley A. HARTLAUB, Angela M. DEAN and Douglas A. WOLFE Rank-based test procedures for interaction in the two-way layout with one observation per cell	863
Oswaldo MARRERO L'analyse de la variation saisonnière quand l'amplitude et la taille sont faibles	875
Index: Volume 27 (1999)	883
Forthcoming Papers/Articles à paraître	890
Volume 28 (2000): Subscription rates/Frais d'abonnements	892

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue of *Survey Methodology* (Vol. 19, No. 1 and onward) as a guide and note particularly the points below. Accepted articles must be submitted in machine-readable form, preferably in WordPerfect. Other word processors are acceptable, but these also require paper copies for formulas and figures.

1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size ($8\frac{1}{2} \times 11$ inch), one side only, entirely double spaced with margins of at least $1\frac{1}{2}$ inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, etc.
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (e.g., w, ω ; o, O, 0; l, 1).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, priez d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 19, n° 1) et de noter les points ci-dessous. Les articles acceptés doivent être soumis sous forme de fichiers de traitement de texte, préféablement WordPerfect. Les autres logiciels sont acceptables, mais une version sur papier sera alors exigée pour le traitement des formules et des figures.

1. Présentation

- 1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.
- 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
- 1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
- 1.4 Les remerciements doivent paraître à la fin du texte.
- 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. Résumé

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. Rédaction

- 3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
- 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(-) et log(-) etc.
- 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
- 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
- 3.5 Distinguer clairement les caractères ambigus (comme w, ω; o, O; l, 1).
- 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.

4. Figures et tableaux

- 4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
- 4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois).

5. Bibliographie

- 5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
- Exemple: Cochran (1977, p. 164).
- 5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

671	Statistical science: concepts, opportunities and challenges	Jerald F. LAWLESS
683	Dirichlet forms: some infinite-dimensional examples	Byron SCHMIDLAND
701	Bayesian variable selection for proportional hazards models	Joseph G. BRAHIM, Ming-Hui CHEN and Steven N. MacEachern
719	Semiparametric Bayesian inference for regression models	Yodit SEIFU, Thomas A. SEVERINI and Martin A. TANNER
735	A generalized-moments specification test for the logistic link	Konstantinos FOKIANOS, Amy PENG and Jing QIN
751	Robust extrapolation designs and weights for biased regression models with heteroscedastic errors	Zhide FANG and Douglas P. WIENS
771	Nonrobustness of the information test in detecting heterogeneity	Michael P. JONES
781	Minimax designs for approximately linear models with AR (1) errors	Douglas P. WIENS and Julie ZHOU
795	Deux méthodes d'estimation pour les paramètres de processus moyenne mobile spatiaux	Luc D. ADJENGUE and Marc MOORE
819	Conditional log-spline density estimation	Benoît R. MASSÉ and Young K. TRUONG
833	Fisher information in weighted distributions	Satish IYENGAR, Paul KVAM and Harshinder SINGH
843	A cluster problem as defined by nearest neighbours	E.G. ENNS, P.F. EHLERS and T. MISI
853	Lois limites pour les statistiques d'ordre dans le cas non identiquement distribué	Mohammadi BELBACHIR
863	Rank-based test procedures for interaction in the two-way layout with one observation per cell	Bradley A. HARTLAUB, Angela M. DEAN and Douglas A. WOLFE
875	L'analyse de la variation saisonnière quand l'amplitude et la taille sont faibles	Oswaldo MARRERO
883	Index: Volume 27 (1999)	
890	Forthcoming Papers/Articles à paraître	
892	Volume 28 (2000): Subscription rates/Frais d'abonnements	

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board

Contents Volume 15, Number 4, 1999

Bayesian Estimation of the Number of Unseen Studies in a Meta-Analysis <i>Lynn E. Eberly and George Casella</i>	477
Toward a Social Psychological Programme for Improving Focus Group Methods of Developing Questionnaires <i>Katherine Bischooping and Jennifer Dykema</i>	495
Statistical Methods for Developing Ratio Edit Tolerances for Economic Data <i>Katherine Jenny Thompson and Richard S. Sigman</i>	517
A Conditional Analysis of Some Small Area Estimators in Two Stage Sampling <i>Piero D. Falorsi and Aldo Russo</i>	537
Internal Migration: What Data are Available in Europe? <i>Philip Rees and Marek Kupiszewski</i>	551
A Bibliography on Statistical Consulting and Training <i>Hardeo Sahai and Anwer Khurshid</i>	587
Editorial Collaborators	631
Index to Volume 15, 1999	663

All inquiries about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Department, Statistics Sweden, Box 24 300, S - 104 51 Stockholm, Sweden.

un plus petit écart quadratique prévu relativement à X_N que (20) et (19) même en l'absence de la restriction supplémentaire notée au paragraphe précédent.

9. NOISE/TONO

Puisque les conditions dans lesquelles l'estimateur (5) de β est plus efficace que l'estimateur (4) sont très restrictives,

Soit $C'(X)Y$, un estimateur arbitraire inconditionnelle-
ment non biaisé de β , où $C(X)$ est une matrice de fonctions
de X , de dimensions identiques à celles de X . L'exigence
d'une absence inconditionnelle de biais suppose la
restriction $E[C'(X)X] = I$ (Shaffer 1991). Conditionnant
d'abord en X et utilisant l'expression pour la variance
inconditionnelle, il s'ensuit que la variance
de $C'(X)Y$ est $E[C'(X)C(X)]\sigma^2 + \text{Var}(C'(X)X\beta)$. Puisque
nous considérons la variance en $\beta = 0$, seul le premier
nombre est non nul. À supposer que $C'(X) =$
 $[E(X'X)]^{-1}X'$, la variance de $\hat{\beta}^*$ est $[E(X'X)]^{-1}\sigma^2$.

Soit $\hat{\beta}$, un estimateur arbitraire inconditionnellement non
biaisé de forme $C'(X)Y$. Des lors, $\text{Var}(\hat{\beta}) = \text{Var}(C'(X)Y) +$
 $\text{Var}(\hat{\beta} - \hat{\beta}^*) + 2\text{Cov}(\hat{\beta}^*, \hat{\beta} - \hat{\beta}^*)$, de sorte que $\text{Var}(\hat{\beta}) <$
 $\text{Var}(\hat{\beta}^*)$ si $\text{Cov}(\hat{\beta}^*, \hat{\beta} - \hat{\beta}^*) > 0$, ou si $\text{Cov}(\hat{\beta}^*, \hat{\beta}) \geq$
 $\text{Var}(\hat{\beta}^*)$. Un calcul facile, à l'aide de la restriction
 $E[C'(X)X] = I$, permet de montrer que $\text{Cov}(\hat{\beta}^*, \hat{\beta}) =$
 $\text{Var}(\hat{\beta}^*)$, ce qui prouve que $\hat{\beta}^{(0)}$ est LBLUE en β_0 .

ANNEXE B

Preuve du résultat de la section 4

$$\begin{aligned} & {}^0\mathfrak{g}^u\mathbf{X} - {}^u\mathbf{X} + {}^0\mathfrak{g}^N\mathbf{X} = \\ & ({}^0\mathfrak{g}^u\mathbf{X} - {}^u\mathbf{X})\mathbf{I}(u/\mathbf{I}) + {}^0\mathfrak{g}^N\mathbf{X} = \\ & ({}^0\mathfrak{g}^u\mathbf{X} - {}^u\mathbf{X})\cdot {}^u\mathfrak{g}^u\mathbf{X} \\ & \mathbf{I}\cdot ({}^N\mathbf{X}\mathbf{I}\cdot {}^N\mathfrak{g}^N\mathbf{X})^N\mathbf{X}\mathbf{I}\cdot {}^N\mathfrak{g}^N\mathbf{X}\mathcal{S}(u/\mathbf{I}) + {}^0\mathfrak{g}^N\mathbf{X} = \\ & ({}^0\mathfrak{g}^u\mathbf{X} - {}^u\mathbf{X})\cdot {}^u\mathfrak{g}^u\mathbf{X} \\ & [\mathbf{I}\cdot ({}^N\mathbf{X}\mathbf{I}\cdot {}^N\mathfrak{g}^N\mathbf{X})(N/\mathbf{I})u]^N\mathbf{X}\mathbf{I}(N/\mathbf{I}) + {}^0\mathfrak{g}^N\mathbf{X} = ({}^0\mathfrak{g})\cdot {}^N\mathfrak{g}^N\mathbf{X} \end{aligned}$$

où B_N et B_n sont les matrices appropriées de la population et de l'échantillon, respectivement. L'expression finale en (B.1) est l'estimateur par la différence généralisé fondé sur une valeur β_0 choisie indépendamment de l'échantillon. Cela prouve la partie (b) du résultat; puisque l'estimateur par la différence est sans biais pour X dans un échantillon autopondéré, le résultat en (a) s'ensuit.

BIBLIOGRAPHIE

LEHMANN, E.L. (1966). Some concepts of dependence. *Annals of Mathematical Statistics*, 37, 1137-1153.

ROYAL, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.

SÄRNÄDAL, C.-E., SWENSSON, B., et WRETMAN, J. (1991). *Model Assisted Survey Sampling*. New York: Springer.

SHAFER, J.P. (1991). The Gauss-Markov theorem and random regressors. *The American Statistician*, 45, 269-273.

des caractéristiques de la population fondées sur (5), il est possible que les résultats présentés ici revèlent un intérêt plus théorique que pratique. Les résultats permettent cependant de mieux comprendre certaines situations dans lesquelles des estimateurs simples comme la moyenne de l'échantillon et l'estimation générale sont efficaces comme moyen d'estimer la moyenne de la population que ne le sont les estimateurs par le quotient, les estimateurs de stratification a posteriori, les estimateurs de régression et d'autres estimateurs complexes. Les équations (6) et (7) pour des variances comparatives de (4) et de (5) fournissent une autre façon de comparer des variances respectives pour différents modèles de régression et différents valeurs de β . Plusieurs de ces résultats sont valables pour des plans d'échantillonnage très simples, mais il devrait être possible de les généraliser sous forme de plans d'échantillonnage à probabilités inégales plus complexes.

L'auteur tient à remercier Erik N. Torgersen, décédé, qui a suggéré l'estimateur généralisé à propriétés optimales, de même que Phillip S. Kott, dont la suggestion a permis de définir la condition d'absence de biais dû au plan d'échantillonnage, ainsi que des examinateurs anonymes de leurs nombreuses et précieuses remarques.

REMERCIEMENTS

Supposons un modèle (1), avec $\text{Var}(Y|X) = \sigma^2 B$. (La preuve générale présentée ici s'applique directement au modèle (10) lui aussi.) Prenons l'estimateur

ANNEXE A

Preuve que $\hat{\beta}_{\beta_0}^*$ est un estimateur BLUE en β_0

$$\hat{p}_*^{(p)} = p_0 + [E(X'B^{-1}X)]^{-1}X'B^{-1}(Y - Xp_0).$$

Supposons $\tau = \beta - \beta_0$ et $Z = Y - X\beta_0$. Dès lors $E(Z|X) = X\tau$, $\text{Var}(Z|X) = \sigma^2 B$, et $\frac{\tau}{\sigma} = [E(X' B^{-1} X)]^{-1} X' B^{-1} Z = \beta^{(B)}$.

permettant l'ajout d'une constante, est l'estimateur LBLUE en $\beta = \beta_0$, pour un vecteur arbitraire β_0 . La preuve de ces résultats se trouve à l'annexe A. Cet estimateur généralisé (13) peut être utile dans une situation d'échantillonnage rapproché d'une valeur donnée. On peut facilement montrer que la variance de (13) est égale à (7) avec $(\beta - \beta_0)$ substitué à β . (Voir l'annexe A.) Dans le cadre du modèle (10), les estimateurs (8), (9) et (12) se laissent généraliser sous la forme

$$\hat{\beta}_{(\beta_0, N)}^* = \beta_0 + [E(X_N' B_N^{-1} X_N)]^{-1} [X_N' B_N^{-1} (Y_N - X_N \beta_0)], \quad (14)$$

$$\hat{\beta}_{(\beta_0, n)}^* = \beta_0 + [E(X_n' B_n^{-1} X_n)]^{-1} [X_n' B_n^{-1} (Y_n - X_n \beta_0)], \quad (15)$$

$$\hat{\beta}_{(\beta_0, n)}^* = \beta_0 + [(n/N) X_N' B_N^{-1} X_N]^{-1} [X_n' B_n^{-1} (Y_n - X_n \beta_0)], \quad (16)$$

respectivement.

4. CONDITIONS DE L'ABSENCE DE BIAIS DÙ AU PLAN D'ÉCHANTILLONNAGE

Nous supposons que le modèle (10) est valable et que l'estimateur non biaisé sans condition peut s'exprimer sous la forme (16). Nous supposons également qu'il existe un échantillon de taille n , $B_n^{-1} X_n' g = 1_N$ et que, pour tout vecteur $p \times 1$ g tel que $B_n^{-1} X_n' g = 1_N$ et 1_n , sont des vecteurs de un de longueur N et n , respectivement. Alors, pour un échantillon aléatoire simple,

$$\begin{aligned} \text{a) l'estimateur} \\ Y_{(\beta_0, n)}^* &= X_n' \hat{\beta}_{(\beta_0, n)}^* \\ (17) \end{aligned}$$

$$\begin{aligned} \text{b) } Y_{(\beta_0, n)}^* &\text{ est un estimateur par la différence généralisé} \\ &\text{est un estimateur sans biais dû au plan d'échan-} \\ &\text{tillonnage de } Y_N, \text{ où } X_N' = (1/N) 1_N' X_N, \text{ et} \\ &\text{de } Y_N. \end{aligned}$$

La preuve figure à l'annexe B.

À noter qu'il existe un vecteur g satisfaisant les conditions de ce théorème si le modèle englobe une ordonnée à l'origine (X_N englobe une colonne de un) ou si B_N est diagonale et la variance est proportionnelle aux valeurs de l'une des variables explicatives. De nombreuses applications de la modélisation de régression à l'estimation par sondage se fondent sur des modèles qui englobent ces hypothèses. Särndal, Swensson and Wretman (1991, p. 231

et 232) aborde ces modèles et d'autres modèles plus généraux, et le chapitre 6 de la section 4 de cet ouvrage offre des exemples de modèles répandus qui intègrent ces hypothèses. Le chapitre 6 comme tel aborde à la fois l'estimateur par la différence généralisé de Y_N et l'estimateur de régression général analogue fondé sur β_n . Ce chapitre propose également des généralisations de ces résultats pour des estimateurs et des plans de sondage plus complexes.

5. DISCUSSION

Afin d'appliquer les résultats à des estimations des propriétés d'une population finie, nous supposons que la matrice B est diagonale ou qu'elle possède la forme diagonale par blocs spéciale et le plan de sondage connexe décrit ci-dessus. D'après les résultats de la section 3, il s'ensuit que l'estimateur (17) de Y_N comporte une plus petite variance que l'estimateur

$$\begin{aligned} \text{(18)} \quad Y_{(\beta_n)}^* &= X_n' \hat{\beta}_n^* \\ &\text{lorsque } \beta \text{ est proche de } \beta_0. \text{ À noter que (18) peut s'écrire} \\ &\text{sous la forme} \end{aligned}$$

$$\text{(19)} \quad Y_{\beta_n}^* = \frac{1}{N} \left[\sum_{i \in s} X_i' \hat{\beta}_n^* + \sum_{i \notin s} X_i' \hat{\beta}_n^* \right],$$

que X_i' est la i -ième ligne de X , et que S est l'ensemble des éléments de l'échantillon. Royall (1970) a montré que le meilleur estimateur linéaire sans biais dû au modèle de Y_N (sous condition pour l'échantillon obtenu) est

$$\text{(20)} \quad \frac{1}{N} \left[\sum_{i \in s} Y_i + \sum_{i \notin s} X_i' \hat{\beta}_n^* \right].$$

Dans certains cas importants, le premier membre de (20) est égal au premier membre de (19), et alors (20) et (19) sont identiques. Ce sera le cas, par exemple, si $B = \sigma^2 I$ et si le modèle (10) comporte une ordonnée à l'origine, ou si $p = 1$ et B est diagonale avec entrées diagonales proportionnelles aux valeurs de la variable explicative unique. Dans de tels cas, (20) et (19) sont identiques, et l'estimateur (17) sans biais dû au plan d'échantillonnage et inconditionnellement quadratique prévu plus petit que (20) lorsque β est proche de β_0 , même en l'absence de l'exigence que les premiers membres de (20) et de (19) soient égaux. Si l'on remplace β par $\hat{\beta}^*$ en (20), l'estimateur obtenu n'est plus inconditionnellement non biaisé. Il est possible de montrer, toutefois, à l'aide de concepts de dépendance (Lehmann, 1966) que, pour les conditions en B notées au début de la présente section, l'estimateur obtenu comportera

variance que lorsque β est petit. En réalité, lorsque $E(X'X)$ est connu, il n'existe aucun meilleur estimateur linéaire sans

biais.

Une comparaison des variances de (2) et de (3) en fonction de diverses hypothèses de modélisation, à part les répercussions pour une estimation des coefficients eux-mêmes, permet de mieux comprendre les conditions dans lesquelles divers estimateurs d'autres paramètres de la population ont des propriétés souhaitables, fondées à la fois sur un modèle et sur le plan de sondage.

2. GÉNÉRALISATION DE LA MATRICE DES COVARIANCES DE ϵ

Si la matrice des covariances de ϵ est de forme $\sigma^2 B$, où B est une matrice définie positive fixe et connue, le théorème de Gauss-Markov s'applique à l'estimateur généralisé

$$(4) \quad \hat{\beta} = [X'B^{-1}X]^{-1}X'B^{-1}Y.$$

Les preuves présentées dans Shaffer (1991) se laissent généraliser directement pour montrer que, si $E(X'B^{-1}X)$ est connu, l'estimateur

$$(5) \quad \hat{\beta}^* = [E(X'B^{-1}X)]^{-1}X'B^{-1}Y$$

comporte une variance plus petite que (4) lorsque β est suffisamment proche de zéro. Les variances (sans condition) de (4) et de (5) sont

$$(6) \quad \Sigma_{\hat{\beta}} = E[X'(X'B^{-1}X)^{-1}] \sigma^2$$

et

$$(7) \quad \Sigma_{\hat{\beta}^*} = [E(X'B^{-1}X)]^{-1} \sigma^2 + \text{Var}\{[E(X'B^{-1}X)]^{-1}(X'B^{-1}X)\beta\}.$$

Lorsque $\beta = 0$, Shaffer montre que (7) est plus petit que (6), et, par conséquent, en supposant la continuité de (7) en fonction de β , plus petit que (6) lorsque β se rapproche de zéro.

Les résultats sont maintenant appliqués dans le contexte de l'échantillonnage. Soit X_N la matrice $N \times p$ et Y_N le vecteur $N \times 1$, dans une population finie. Si les éléments N de la population sont considérés comme un échantillon tiré d'une population hypothétique infinie d'éléments potentiels satisfaisant (1), et si l'on tire un échantillon de taille n de la population finie, les preuves présentées dans Shaffer (1991) se laissent généraliser directement pour montrer que

$$(8) \quad \hat{\beta}_N = [E(X_N'B_N^{-1}X_N)]^{-1}X_N'B_N^{-1}Y_N$$

et que

$$(9) \quad \hat{\beta}_n^* = [E(X_n''B_n''^{-1}X_n'')]^{-1}X_n''B_n''^{-1}Y_n''$$

comportent des variances plus petites que celles de leurs versions correspondantes sous condition $\hat{\beta}_N$ et $\hat{\beta}_n$ respectivement, si β se rapproche de zéro, où l'espérance en (8) se rapporte à la population infinie d'éléments hypothétiques, et l'espérance en (9) se rapporte à la même population finie d'éléments N satisfaisant (1). Si l'on veut appliquer ces résultats, il faut que les espérances en (8) et en (9) soient connues. Si l'on considère X_N comme fixe, le modèle de la population se laisse écrire sous la forme

$$(10) \quad Y_N = X_N\beta + \epsilon_N,$$

où ϵ_N est un vecteur de termes d'erreur distribués au hasard comme en (1). Dans le cadre du modèle (10), $\hat{\beta}_N$ et $\hat{\beta}_n$ sont identiques, mais $\hat{\beta}_n$ demeure distinct de $\hat{\beta}_n^*$. Dans le cadre du modèle (10), pour un échantillon aléatoire de taille n , si

$$(11) \quad E[(X_n''B_n''^{-1}X_n'')/n] = (X_N'B_N^{-1}X_N)/N,$$

l'autre estimateur peut s'écrire sous la forme

$$(12) \quad \hat{\beta}_n^* = [(n/N)(X_N'B_N^{-1}X_N)]^{-1}X_n''B_n''^{-1}Y_n''.$$

Dans le modèle (10), les équations (11) et (12) s'appliquent si B_N est diagonale et si le plan d'échantillonnage est autopondéré, et pour certaines autres conditions et plans d'échantillonnage, par exemple si B_N est diagonale par blocs (grappes) et des grappes complètes sont échantillonnées. Si B_N est diagonale, B_n n'est pas nécessairement fixe. Supposons, par exemple, une population constituée d'hommes et de femmes, les variances des deux sexes pour la caractéristique d'intérêt étant connues et différentes. Dans un tel cas, si l'on tire un échantillon autopondéré, et si l'on suppose que le modèle (10) s'applique aux deux sous-populations, B_n sera diagonale, les entrées étant fonction des proportions des deux sexes dans l'échantillon.

3. MEILLEURE ESTIMATION LINÉAIRE SANS BIAIS LOCALEMENT

Dans le cadre du modèle (1), l'estimateur (5) est le meilleur estimateur linéaire sans biais localement (LBLUE) lorsque $\beta = 0$; c'est-à-dire l'estimateur linéaire en X et sans biais pour β avec la plus petite variance dans un voisinage de $\beta = 0$. De plus, l'estimateur linéaire généralisé

$$(13) \quad \hat{\beta}^{(\beta_0)} = \beta_0 + [E(X'B^{-1}X)]^{-1}[X'B^{-1}(Y - X\beta_0)],$$

Les meilleurs estimateurs linéaires sans biais locaux et non conditionnels: applications à l'échantillonnage

JULIET POPPER SHAFER¹

RÉSUMÉ

Les statisticiens d'enquête ont fréquemment recours à des modèles de régression linéaire de superpopulation. Le théorème de Gauss-Markov, qui suppose des variables explicatives fixes ou un conditionnement des valeurs observées de celles-ci, affirme que les estimateurs standard des coefficients de régression sont les meilleurs estimateurs linéaires sans biais. Shaffer (1991) a montré que le théorème de Gauss-Markov ne s'applique pas lorsque les variables explicatives sont aléatoires si certains aspects de la distribution de la population des variables explicatives sont connus, et a présenté un autre estimateur ayant de meilleures propriétés que l'estimateur standard sous certaines conditions. Le présent exposé établit certaines généralisations et constate pour cet autre estimateur généralisé une absence de biais meilleure localement (propriété d'optimalité). L'auteur décrit les répercussions pour l'étape d'estimation des enquêtes.

MOTS CLÉS: Analyse de régression; théorème de Gauss-Markov; échantillonnage; estimation sans biais; optimalité; meilleure estimation linéaire sans biais.

1. INTRODUCTION

Dans le modèle standard de régression linéaire pour un échantillon d'observations,

$$Y = X\beta + \epsilon, \quad (1)$$

Il s'agit d'abord de présenter certains résultats relevant du modèle général (1). Ensuite, il est question des modifications qui s'appliquent à l'échantillonnage. En vertu du modèle (1) avec $\sum = \sigma^2 I$, le théorème de Gauss-Markov affirme que l'estimateur

$$\hat{\beta} = (X'X)^{-1}X'Y, \quad (2)$$

est un meilleur estimateur linéaire sans biais si X est considérée comme une matrice fixe. Si les lignes de X sont traitées comme des réalisations de vecteurs aléatoires $x_i, i = 1, \dots, n$, le théorème de Gauss-Markov peut être interprété comme une déclaration voulant que l'estimateur en (2) comporte une variance minimale dans la catégorie d'estimateurs linéaires en X et sans biais sous condition, étant donné ces valeurs réalisées de X . Toutefois, l'utilisation du terme «sans biais» sans autres précisions signifie généralement une absence de biais sans condition. Si l'exigence d'absence de biais est interprétée comme une absence de biais sans condition, c'est-à-dire en moyenne pour des vecteurs aléatoires ayant des valeurs en X , le théorème de Gauss-Markov, comme l'a indiqué Shaffer (1991), ne s'applique pas lorsque $E(X'X)$ est connu. Dans ce cas, l'estimateur biaisé sous condition

$$\hat{\beta}^* = [E(X'X)]^{-1}X'Y \quad (3)$$

a) Les résultats sont généralisés à partir d'un modèle dans lequel la matrice des covariances de l'échantillon des erreurs ϵ est $\sigma^2 I$, où I est la matrice d'identité $n \times n$, pour le cas dans lequel la matrice des covariances Σ de ϵ est $\sigma^2 B$, où B est une matrice définie positive fixe et connue, et pour certaines situations dans lesquelles B est aléatoire (puisque il s'agit de la matrice des covariances d'un échantillon choisi au hasard des valeurs des variables explicatives).

b) On obtient un estimateur généralisé qui fonctionne bien lorsque le vecteur coefficient β se rapproche de tout vecteur coefficient préétabli β_0 .

¹ Juliet Popper Shaffer, Department of Statistics, 367 Evans Hall, #3860, Berkeley, CA 94720-3860, U.S.A.
Courriel: shaffer@berkeley.edu

REMERCIEMENTS

Les auteurs sont reconnaissants envers le National Agricultural Statistics Service qui a, au départ, sélectionné l'échantillon et les enquêteurs du National Agricultural Statistics Service qui ont, les premiers, contacté les fermes, les vétérinaires et les techniciens vétérinaires du gouvernement fédéral et des gouvernements étatiques qui ont effectué les autres visites sur les lieux, et tous les éleveurs

BIBLIOGRAPHIE

- de bovins de boucherie admissibles, qu'ils aient ou non participé à l'enquête.
- RIZZO, L., KALTON, G., et BRICK, J. M. (1996). Comparaison de quelques méthodes de correction de la non-réponse d'un panel. *Techniques d'enquête*, 22, 43-53.

non-réponse de l'unité, nous recommandons que les statistiques d'enquête respectent d'abord certaines procédures afin de déterminer dans quelle mesure les variables sont liées à d'autres caractéristiques des unités éligibles l'enquête. La correction des poids en fonction uniquement des variables qui se sont avérées de bons indices de la non-réponse du panel peut donner lieu à des estimations de la population faussées si les variables ne constituent pas aussi de bons prédicteurs des données que les non-répondants auraient fournies à l'enquête.

Tableau 2

Variables sélectionnées à partir des deux premières étapes de la collecte des données						
Variables auxiliaires						
enviagées pour la correction						
à la troisième étape						
Nombre de vaches d'élevage						
de bouchère						
1 - 49						
50 - 99						
100+						
69,2	50,8	63,1	15,4	215	6,3	
69,2	59,6	80,8	26,9	232	3,9	
88,2	70,1	85,4	52,8	223	4,1	
Nombre de périodes de reproduction						
1						
>1 ou aucune période de reproduction définie						
62,3	69,8	17,0	223	5,1		
-	63,5	81,3	43,8	223	4,5	
On a consulté un vétérinaire pour traiter ou diagnostiquer une maladie en 1996						
79,2	-	69,8	222	4,5		
Non	80,0	-	84,2	223	4,6	

Variables sélectionnées à partir des deux premières phases de la collecte des données:

- 1 = Opérations comptant une période de reproduction définie
- 2 = Opérations ayant consulté un vétérinaire pour traiter ou diagnostiquer une maladie en 1996
- 3 = Opérations ayant fait vacciner leurs génisses contre la brucellose
- 4 = Opérations ayant implanté un stimulateur de croissance dans leurs veaux avant ou au moment du sevrage en 1996
- 5 = Âge moyen (en jours) des veaux au moment du sevrage
- 6 = Pourcentage de veaux morts en 1996

vétérinaire en 1996. Cependant, le pourcentage des opérations ayant consulté un vétérinaire était presque identique selon qu'il s'agissait d'opérations comptant une période de reproduction définie ou d'opérations ne comptant pas de période de reproduction définie. De plus, le pourcentage des opérations ayant fait vacciner leurs génisses contre la brucellose et le pourcentage des opérations ayant fait implanter à leurs veaux un stimulateur de croissance ont affiché une gamme plus grande selon la taille de troupeau que selon les deux autres variables auxiliaires enviagées. En outre, l'âge moyen de sevrage et la perte moyenne de veaux ont davantage varié selon la taille du troupeau que selon le nombre de périodes de reproduction ou selon que l'on a consulté ou non un vétérinaire. On a observé des modèles semblables dans les autres régions.

Bien que la taille du troupeau n'ait pas été un indice statistiquement significatif de la participation à la dernière étape de la collecte des données dans le cadre de l'étude sur les bovins de 1997 du NAHMS (tableau 1), on a constaté que la taille du troupeau se rapportait davantage à un certain nombre de variables du questionnaire qu'à l'une des autres variables auxiliaires enviagées tirées de l'analyse de régression logistique. Par conséquent, nous avons utilisé la région traditionnelle d'après le modèle de la catégorie de la taille du troupeau pour corriger les poids selon la non-réponse à la dernière étape de la collecte des données dans le cadre de l'étude NAHMS.

Les chercheurs qui s'appuient sur les données d'enquête dépendent des poids d'échantillonnage pour produire des estimations des paramètres de population qui sont approximativement non biaisées. Dans le cadre de l'étude sur les bovins de 1997 du NAHMS, une analyse de régression logistique menée lors de la dernière étape de collecte des données a permis de définir deux variables plus performante que la taille du troupeau pour prédire la non-réponse à la dernière étape de la collecte des données. Cependant, l'efficacité de ces variables s'est avérée en général inférieure à celle obtenue pour la taille du troupeau pour ce qui est d'établir des distinctions quant à la façon dont les éleveurs ont répondu à un certain nombre de questions clés concernant la gestion des opérations. Si l'on s'était servi de ces deux variables pour établir des catégories en vue de la correction des poids selon la non-réponse, on aurait réduit le biais dans les estimations des paramètres (de la troisième étape de la collecte des données) avec lesquels elles sont en corrélation. Cependant, les estimations des paramètres qui n'étaient pas en corrélation avec ces variables auraient pu être biaisées. Par conséquent, nous avons choisi l'approche conventionnelle pour corriger les poids selon la non-réponse d'après la région et la taille du troupeau.

La définition de variables qui sont de bons indices de la non-réponse du panel constitue une bonne pratique dans le cadre de n'importe quelle enquête à plusieurs étapes. Avant de se servir de ces variables pour corriger les poids selon la

reproduction définie, et de 1,875 pour les opérations qui comportaient une période de reproduction définie. Pour ce qui est du troisième modèle, les opérations dans l'Ouest qui compartaient une période de reproduction définie étaient divisées en deux cellules selon qu'elles avaient ou non reconnu aux services d'un vétérinaire pour diagnostiquer ou traiter une maladie durant 1996: les opérations ayant répondu «oui» ont reçu une correction de poids de 1,548, tandis que les opérations ayant répondu «non» ont reçu une correction de poids de 2,326.

Tableau 1

Les résultats de la régression logistique progressive visant à définir les variables associées à la non-réponse à la dernière étape de la collecte des données dans le cadre de l'étude sur les bovins de 1997 du National Animal Health Monitoring System. D'après 1 190 opérations admissibles et 238 non-répondants

Variable/ Réponses	Estimation des paramètres	p
Coordonnée à l'origine	0,369	0,181
Région	0,851	0,000
Centre-Nord	0,822	0,000
Centre-Sud	2,062	0,000
Centre	1,164	0,000
Sud-Est	1,000	
Ouest	0,299	0,106
Nombre de vaches d'élevage	1 - 49	
de boucherie	50 - 99	0,146
	100 +	1,000
Nombre de périodes de reproduction	1	-0,370
	> 1 ou aucune période établie	1,000
On a consulté un vétérinaire pour traiter ou diagnostiquer une maladie en 1996.	Oui	0,441
	Non	1,000

Pour étudier dans quelle mesure les variables auxiliaires envisagées se rapportaient aux stratégies de gestion globales, nous avons sélectionné d'autres variables à partir des deux premières étapes de la collecte des données et, dans chaque région, nous avons étudié les différences dans ces variables selon la taille du troupeau, selon le nombre de périodes de reproduction et selon que l'on a consulté ou non un vétérinaire pour diagnostiquer ou traiter une maladie en 1996. Le tableau 2 présente certains résultats représentatifs pour la région de l'Ouest. Il y avait des différences dans la taille du troupeau dans le pourcentage des opérations qui avaient consulté un

taux de réponse à la dernière étape de 45 variables établies en fonction des données recueillies durant les deux premières étapes d'interviews. On s'est servi d'une procédure de sélection des variables progressive, dans le cadre de laquelle les régions et les tailles du troupeau s'inscrivaient obligatoirement dans un modèle de régression logistique et dont le niveau de signification était de 0,05 pour l'entrée et le maintien des autres variables dans le modèle (tableau 1). L'analyse de la régression logistique a montré que la réponse à la dernière étape différait quelque peu selon la région et de façon négligeable selon la taille du troupeau. On a associé l'augmentation de la non-réponse à l'existence d'une seule période de reproduction et au fait de ne pas avoir consulté le vétérinaire pour le traitement ou le diagnostic de maladies en 1996. On a étudié la possibilité de se servir de variables de régression logistique comme variables auxiliaires pour la création de cellules permettant la correction des poids selon la non-réponse à la dernière étape. On a proposé quatre modèles de catégorisation:

1. La région conventionnelle selon la taille du troupeau comptant 15 cellules.
2. La région selon la taille du troupeau, sauf dans l'Ouest, où elle a été subdivisée par le nombre de périodes de reproduction, pour un total de 14 cellules.
3. La subdivision des cellules de l'option 2 (par l'un des variables auxiliaires) si la différence dans le taux de réponse (entre les deux nouvelles subdivisions) était d'au moins 10% et qu'au moins 20 répondants demeuraient dans chaque cellule. On a procédé à deux subdivisions, qui ont donné lieu à un total de 16 cellules.
4. La poursuite de la subdivision des catégories, en fonction du plus grand écart dans le taux de réponse, jusqu'à ce qu'un nombre minimal de répondants (pas moins de 20) demeurent dans chaque cellule. Cela a donné lieu à un total de 24 cellules.

On a calculé les facteurs de correction des poids des répondants à la dernière étape en divisant la somme des poids de la deuxième étape qui s'appliquent aux opérations admissibles par la somme des poids de la deuxième étape qui s'appliquent aux répondants de la dernière étape dans

chaque cellule. Comme l'établissement des cellules pour les modèles 2 à 4 s'est fait selon les variables ayant démontré les plus grands écarts dans le taux de réponse, les différences dans les facteurs de correction ont augmenté à l'égard de sous-catégories particulières du modèle 1 au modèle 4. Par exemple, pour le premier modèle, les facteurs de correction pour la région de l'Ouest étaient de 1,897, 1,504 et 1,579 pour les tailles de troupeau petites, moyennes et grandes, respectivement. Pour le deuxième modèle, les facteurs de correction pour la région de l'Ouest étaient de 1,334 pour les opérations qui ne comptaient pas de période de

Mise en garde sur la correction des poids selon la non-réponse

WILLARD C. LOSINGER, LINDSEY P. GARBER, BRUCE A. WAGNER et GEORGE W. HILL¹

RÉSUMÉ

Pour les enquêtes dont la collecte des données comprend plus d'une étape, on recommande, comme méthode de correction des étapes antérieures de la collecte des données (qui sont reconnues comme des prédicteurs de la non-réponse. Dans le cadre de l'étude sur les bovins de 1997 du National Animal Health Monitoring System des États-Unis, on a défini, à la dernière étape de la collecte des données, deux variables qui séparaient clairement les éleveurs admissibles selon leur propension à répondre. Cependant, ces variables étaient nettement moins fiables que les catégories simples de la région d'élevage de la collecte des données, même si d'autres variables étaient de meilleurs indices de réponse. Lors de la sélection des variables auxiliaires qui permettent de corriger les poids selon la non-réponse, nous recommandons que les statistiques d'enquête évaluent aussi dans quelle mesure ces variables auxiliaires se rapportent aux données que les non-répondants auraient fournies. L'utilisation de variables auxiliaires qui affichent le plus grand écart quant à la propension de répondre peut entraîner le plus grand écart quant aux facteurs d'ajustement correcteurs, mais peut biaiser les estimations de la population à l'égard des paramètres non liés aux variables auxiliaires choisies.

MOTS CLÉS : Biases de non-réponse; propension de répondre; régression logistique; enquête nationale.

1. INTRODUCTION

Dans les enquêtes à plusieurs étapes, même si certains participants omettent de répondre à la dernière étape de la collecte des données, on a beaucoup de renseignements sur les non-répondants à la dernière étape d'après les étapes précédentes de l'enquête. Kizzo, Kalton et Brick (1996) ont présenté plusieurs méthodes pour sélectionner les variables auxiliaires et pour corriger les poids selon la non-réponse quand on connaît de nombreuses caractéristiques des non-répondants. Ces méthodes sont axées sur la définition et l'utilisation des caractéristiques qui établissent des distinctions entre les répondants et les non-répondants admissibles. Cependant, en corrigeant les poids en fonction de variables précises qui font état du plus grand écart entre les taux de réponse, on peut biaiser les estimations si ces variables ne sont pas liées aux réponses que les non-répondants auraient données à la dernière étape de la collecte des données. Par conséquent, on doit aussi se servir des données recueillies dans le cadre des étapes précédentes de la collecte des données afin de déterminer si les variables auxiliaires sélectionnées sont liées aux autres caractéristiques des éleveurs qui peuvent participer à l'enquête.

L'étude sur les bovins de 1997 (du National Animal Health Monitoring System (NAHMS) du Department of Agriculture (USDA) des États-Unis) a eu lieu dans 23 États dans le cadre d'une collecte des données en trois étapes. Au février 1997, les enquêteurs du National Agricultural Statistics Service du USDA ont recueilli des données sur

les pratiques de gestion générales auprès de 2 713 opérations agricoles comptant au moins une vache d'élevage de boucherie. Les répondants à la première étape qui comprenaient au moins cinq vaches d'élevage de boucherie le 1^{er} janvier 1997 pouvaient passer à la deuxième étape de la collecte des données (du 3 mars au 23 mai 1997) à condition d'avoir au moins une vache d'élevage de boucherie et d'être toujours en affaire au moment de la deuxième étape de la collecte des données. Au total, 1 190 éleveurs ont participé à la deuxième étape de la collecte des données, qui comprenait une visite sur les lieux effectuée par un représentant de la médecine vétérinaire ou un technicien vétérinaire et portait sur la guidance sanitaire des bovins de boucherie.

Toutes les opérations ayant participé à la deuxième étape de la collecte des données pouvaient prendre part à la troisième et dernière étape de la collecte des données (du 1^{er} août 1997 au 31 janvier 1998). Au total, 952 (80,0 %) opérations admissibles ont participé à la dernière étape. D'après les deux premières étapes de la collecte des données, on disposait de nombreux renseignements sur les 238 non-répondants à la dernière étape. La présente mise en garde vise à décrire les méthodes qui ont été évaluées pour corriger les poids d'échantillonnage de manière à tenir compte de la non-réponse à la dernière étape de la collecte des données dans le cadre de l'étude sur les bovins de 1997 du NAHMS.

En plus d'examiner les catégories de la région et de la taille du troupeau (établie selon le nombre de vaches d'élevage de boucherie), on a évalué l'incidence sur les

- Théberge: Calage et poids restreints
- BREWER, K.R.W. (1979). A class of robust sampling designs for large scale surveys. *Journal of the American Statistical Association*, 74, 911-915.
- DEVILLE, J.-C., et SÄRDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- DUCHESNE, P. (1999). Estimateurs de calage robustes. *Techniques d'enquête*, 25, 47-60.
- FAN, K. (1956). On systems of linear inequalities. *Annals of Mathematics Studies*, (Eds. H. W. Kuhn, et A. W. Tucker), 38, 99-156.
- GRAYBILL, F.A. (1983). *Matrices with Applications in Statistics*. (Deuxième édition). Belmont, California: Wadsworth Publishing.
- RAO, J.N.K., et SINGH, A.C. (1997). A ridge-shrinkage method for range-restricted weight calibration in survey sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.
- SÄRDAL, C.-E., SWENSSON, B., et WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- THEBERGE, A. (1999). Extensions of calibration estimators in survey sampling. *Journal of the American Statistical Association*, 94, 635-644.
- ISAKI, C.T., et FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

moins une des coordonnées est zéro. La fonction Ω varie linéairement sauf à des points Φ perpendiculaires à une ou plusieurs rangées de V . Même lorsque le domaine de Ω est restreint aux vecteurs Φ avec $\|\Phi\|_{l_1} = 1$ qui sont perpendiculaires à $0 < j < (p-1)$ rangées de V linéairement indépendantes, la fonction Ω varie encore linéairement sauf à des points perpendiculaires à d'autres rangées de V ou perpendiculaires à des vecteurs unité (lesquels sont également des rangées de V). Le maximum de Ω pour $\|\Phi\|_{l_1} = 1$ est donc atteint en un point Φ perpendiculaire à $(p-1)$ rangées de V linéairement indépendantes. Il est donc suffisant de vérifier la condition pour deux vecteurs de rangées de V linéairement indépendantes à $(p-1)$ direction opposée qui sont perpendiculaires à $(p-1)$ rangées de V linéairement indépendantes et ce, pour chaque sous-ensemble de $(p-1)$ rangées de V linéairement indépendantes.

ANNEXE B

On note $\text{vec}(F)$ le vecteur obtenu en empilant les colonnes successives de la matrice $F \in \mathbb{R}^{a \times b}$ avec la première colonne au-dessus, et on définit le produit de Kronecker de deux matrices F et G comme

$$F \otimes G = \begin{pmatrix} f_{11}G & \dots & f_{1m}G \\ \vdots & & \vdots \\ f_{m1}G & \dots & f_{mm}G \end{pmatrix}. \quad (\text{B1})$$

Le résultat découle du corollaire de la section 3 avec

$$M = \begin{pmatrix} I_{RC} \\ I^R \otimes I_{1 \times C} \\ I_{1 \times R} \otimes I_C \\ I_{1 \times RC} \end{pmatrix}, \quad w = \text{vec}((N_{ij}^{'})^{'})$$

$$I = \begin{pmatrix} \text{vec}((N_{ij}^{'})^{'}) \\ N_{(T)}^1 \\ \vdots \\ N_{(T)}^R \\ N_{(T)}^1 \\ \vdots \\ N_{(T)}^C \\ N_{(T)}^C \end{pmatrix}, \quad h = \begin{pmatrix} \text{vec}((N_{ij}^{'})^{'}) \\ N_{(H)}^1 \\ \vdots \\ N_{(H)}^R \\ N_{(H)}^1 \\ \vdots \\ N_{(H)}^C \\ N_{(H)}^C \end{pmatrix}. \quad (\text{B2})$$

BIBLIOGRAPHIE

- BACHARACH, M. (1965). Estimating nonnegative matrices from marginal data. *International Economic Review*, 6, 294-310.
- BARDSLEY, P., et CHAMBERS, R.L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33, 290-299.
- BEN-ISRAEL, A., et GREVILLE, T.N.E. (1980). *Generalized Inverses: Theory and Applications*. Huntington, New York: Robert E. Krieger Publishing Company.

donc Φ est perpendiculaire à toutes les lignes de L^* , et à plus forte raison, Φ est perpendiculaire à toutes les lignes de L . De même, le vecteur $\Phi^* = (\Phi_S^*, \Phi_T^*, -1)^*$ est perpendiculaire à toutes les lignes d'un sous-ensemble de $R+C$ lignes de V linéairement indépendantes qui inclut la ligne $RC+i$ ($i = 1, \dots, R$) si et seulement si $i \notin S$, et inclut la ligne $RC+R+j$ ($j = 1, \dots, C$) si et seulement si $j \notin T$, mais qui n'inclut pas la dernière ligne de V . La condition $-1^* \lambda^* \leq h^* \lambda^*$ avec $\lambda^* = V\Phi$ donne le cinquième ensemble d'inéquations dans (9). De même, en posant λ égal à $-V\Phi$, $V\Phi^*$ et $-V\Phi^*$ on obtient les trois derniers ensembles d'inéquations dans (9).

$$V\Phi = ((-\Phi_S \otimes I_{C \times 1} + I_{R \times 1} \otimes \Phi_T^{'})^{'}, \Phi_S^{'}, -\Phi_T^{'}, 0)^*$$

Un sous-ensemble arbitraire de $R+C$ lignes de V linéairement indépendantes qui inclut la dernière ligne de V sera noté L . Si L^* inclut la ligne $RC+i$ ($i = 1, \dots, R$) si et seulement si $i \notin S \subseteq \{1, 2, \dots, R\}$, et inclut la ligne $RC+R+j$ ($j = 1, \dots, C$) si et seulement si $j \notin T \subseteq \{1, 2, \dots, C\}$, alors on pose $\Phi = (\Phi_S^{'}, -\Phi_T^{'}, 0)^*$, où le i -ième élément de $\Phi_S \in \mathbb{R}^R$ est égal à 1 si $i \in S$ et à zéro sinon, et le j -ième élément de $\Phi_T \in \mathbb{R}^C$ est égal à un si $j \in T$ et à zéro sinon. Alors

forment une base pour $N(M^{'})$. C'est-à-dire que $M^{'}/V = 0$, les colonnes de V sont linéairement indépendantes, et $N(M^{'})$ est de dimension $R+C+1$. En notant également que les dernières $R+C+1$ lignes de V sont les vecteurs unité. Et finalement, en vérifiant les conditions du corollaire pour tous les vecteurs $\lambda = V\Phi$ et $\lambda = -V\Phi$, où Φ est perpendiculaire à $R+C$ lignes de V linéairement indépendantes. Cette dernière étape est élaborée plus en détails dans le paragraphe suivant.

$$V = \begin{pmatrix} -I^R \otimes I_{C \times 1} & -I_{R \times 1} \otimes I_C & 0_{R \times R} & 0_{R \times C} & 0_{R \times 1} \\ 0_{C \times R} & I_C & 0_{C \times 1} & 0_{1 \times C} & 1 \end{pmatrix} \quad (\text{B3})$$

Seul un ensemble fini de conditions doit être vérifié, premièrement en notant que les colonnes de

en probabilité vers 0, donc pour des tailles d'échantillons suffisamment grandes, les poids calés w^{cal} demeureront à l'intérieur de la région de restriction R_w si $A_{\mathcal{C}_S}$ est à l'intérieur de R_w . Cependant, on a vu que pour l'estimation d'un total, on n'a pas nécessairement convergence vers 0 de $D_S(w^{\text{cal}})$. Une région de restriction définie par $\|w_S - A_{\mathcal{C}_S}\|_{U_S} \leq l$ est donc à éviter, à tout le moins si on estime un total plutôt qu'un moyen.

On a donné des conditions nécessaires et suffisantes pour l'existence de poids restreints à des intervalles qui satisfont l'équation de calage. Si de tels poids n'existent pas, il faut alors abandonner l'idée de satisfaire exactement l'équation de calage. On peut reformuler le problème de calage avec poids restreints, de manière à ce qu'une solution soit toujours possible. Quelques-unes des approches présentes dans cet article permettent d'obtenir une solution sans avoir recours à des méthodes itératives. Il s'agit de méthodes simples, faciles à interpréter. Les propriétés asymptotiques de ces estimateurs sont habituellement identiques à celles de l'estimateur par calage sans restrictions sur les poids.

Le problème de poids extrêmes survient avec des tailles d'échantillon qui sont petites, donc le problème d'estimation pour de petits domaines devrait être envisagé en même temps. Il est possible de tirer profit d'estimateurs synthétiques tout en ayant un estimateur avec des poids restreints qui a de bonnes propriétés asymptotiques.

On peut aussi modifier l'estimateur par calage, ou tout autre estimateur asymptotiquement convergent, pour traiter les données aberrantes. Les conditions pour que cet estimateur modifié ait les mêmes propriétés asymptotiques que l'estimateur non modifié ne sont pas facilement vérifiables, mais l'estimateur est effectivement aberrante (non représentative). Cependant, ces conditions permettent d'identifier les facteurs qui font qu'un estimateur corrigé pour données aberrantes peut être statistiquement valable.

REMERCIEMENTS

L'auteur tient à remercier un éditeur associé et un arbitre pour leurs remarques constructives qui ont permis d'améliorer cet article.

ANNEXE A

On veut vérifier que $\Omega(\Phi) = l'(V\Phi) - h'(V\Phi)^+$ est inférieur ou égal à zéro. Premièrement, il est facile de montrer que ceci est vrai pour un vecteur Φ , si et seulement si c'est vrai pour un vecteur $k\Phi$ avec $k > 0$ arbitraire. Seulement la direction de Φ importe. Il est donc suffisant de vérifier la condition pour Φ de norme égale à un. Pour la preuve, on utilisera la norme l_1 de Φ , $\|\Phi\|_{l_1} = \sum_{j=1}^p |\phi_j|$. Les vecteurs Φ avec $\|\Phi\|_{l_1} = 1$ sont situés dans des hyperplans dont les intersections sont à des points perpendiculaires aux vecteurs unité, c'est-à-dire des points dont au

raisonnable d'avoir X^{cal} à l'extérieur d'une certaine région. Si X^{cal} est dans la région de restriction (c'est-à-dire si l'expert ne juge pas raisonnable une estimation du paramètre) qui serait égale à la vraie valeur, X^{cal} , du fait que la région de restriction ne varie pas avec n ou N (ou si $\gamma = 1$, et la région de restriction varie proportionnellement à N), alors pour n suffisamment grand, la probabilité que X^{cal} soit à l'intérieur de la région de restriction est égale à un. Dans les cas où X^{cal} est à l'extérieur de la région de restriction, on pourrait utiliser comme estimation le point de la région de restriction qui est le plus près de X^{cal} , ou on pourrait poser égal à un le poids des quelques observations qui sont jugées aberrantes, et répartir leurs poids originaux (moins le nombre d'observations aberrantes) sur les observations non aberrantes. Les propriétés asymptotiques de cet estimateur modifié pour traiter les valeurs aberrantes sont alors les mêmes que celles de l'estimateur non modifié.

Dans le cas d'un échantillon non stratifié cette méthode est relativement facile à appliquer. Si par contre l'échantillon est stratifié, et si des contraintes sont imposées aux estimations de chaque strate, alors on doit faire face à deux problèmes additionnels. Premièrement, si le cadre asymptotique est tel que le nombre de strates augmente de façon proportionnelle à la taille d'échantillon, alors la supposition donnée en (15) ne tient pas, puisque la taille moyenne d'échantillon par strate reste constante alors que $n \rightarrow \infty$. Il s'agit de savoir si l'estimateur de se donner un cadre asymptotique où le nombre de strates est constant (ou croît moins vite que n). Un tel cadre asymptotique est moins plausible si le nombre d'observations par strate est petit. Le deuxième problème est la difficulté pour l'expert d'imposer des contraintes aux estimations pour chacune des strates. Plus il y a de strates, plus le risque est grand que X^{cal} ne soit pas dans la région de restriction définie par l'expert. En fait, dans le cas d'un échantillon stratifié, il est préférable que l'expert utilise les informations indépendantes des données d'enquêtes, afin de s'assurer de l'homogénéité des strates, avant que la stratification soit finalisée. En d'autres mots il vaut mieux utiliser l'information qui est disponible avant l'enquête, pour prévenir les données aberrantes, plutôt que pour les corriger. Si l'information a été utilisée de sorte qu'avant l'enquête, on n'a aucune raison de croire qu'il y a dans quelque strate une observation non représentative, alors on n'a aucune justification pour prétendre le contraire après la collecte des données.

8. CONCLUSION

Si pour de grandes tailles d'échantillons, les poids calés demeurent à l'intérieur d'une région de restriction, alors les propriétés asymptotiques de l'estimateur avec poids restreints sont bien sûr identiques à celles de l'estimateur par calage. Pour un cadre asymptotique donné, on peut habituellement s'attendre à avoir $w^{\text{cal}} - A_{\mathcal{C}_S}$ qui converge

qui est le plus près de w^{cal} . Toujours pour la région de restriction R_w , on aurait

$$w^{\text{res } 4} = \min [\max (w^{\text{cal}}, w^{(L)}), w^{(H)}].$$

Les propriétés asymptotiques des estimateurs qui utilisent les poids restreints $w^{\text{res } 3}$ ou $w^{\text{res } 4}$ sont les mêmes que celles de l'estimateur par calage, pourvu que $w^{\text{cal}} - A^s c_s$ converge en probabilité vers 0, ce qui est normalement le cas.

Une propriété intéressante de toutes les approches en probabilité vers 0, ce qui est normalement le cas.

On veut estimer un total à partir d'un échantillon aléatoire simple de taille 2 d'une population de taille 20. C'est-à-dire $c = 1^{20 \times 1}$ et $a = 10(1^{20 \times 1})$. On utilise le vecteur d'information auxiliaire $X = (1, 2, 3, \dots, 20)'$, on suppose que l'échantillon obtenu est $s = 2, 12$ et on choisit de prendre U une matrice diagonale avec $x_k = k$. On se donne une région de restriction rectangulaire à l'aide des points $w^{(L)} = (0, 0)'$ et $w^{(H)} = (20, 13)'$. C'est-à-dire que le poids de la première unité de l'échantillon doit être supérieur à 0 et inférieur à 20, tandis que le poids de la seconde unité de l'échantillon doit être supérieur à 0 et inférieur à 13.

Sous ces conditions, les poids calés $w^{\text{cal}} = (15, 15)'$ sont à l'extérieur de la région de restriction. Puisque $p = 1$, les poids $w_s(a)$ sont sur le segment de droite qui relie $A^s c_s = (10, 10)'$ à w^{cal} . On a donc $w^{\text{res } 1} = w^{\text{res } 3}$, c'est-à-dire que les deux méthodes donnent le même résultat. Dans ce cas-ci on a $w^{\text{res } 1} = w^{\text{res } 3} = (13, 13)'$. La méthode qui consiste à choisir le point de la région de restriction le plus près des poids calés donne $w^{\text{res } 4} = (15, 13)'$. Par contre, si on cherche $w^{\text{res } 5}$, les poids restreints obtenus en continuant d'exiger que l'équation de calage soit satisfait et en utilisant une mesure de distance qui prend une valeur infinie à l'extérieur de la région de restriction, alors il n'y a pas de solution. En effet, pour tout poids dans la région de restriction $X^s w_s \leq 196$, tandis que $X^s c = 210$. Si on avait, disons $w^{(H)} = (30, 13)'$, alors avec $D^s(w)$ comme mesure de distance à l'intérieur de la région de restriction on aurait $w^{\text{res } 5} = (27, 13)'$. Ces poids sont plutôt éloignés de $w^{\text{cal}} = (15, 15)'$ et de $A^s c_s = (10, 10)'$. C'est le prix qu'il faut payer si on tient à avoir des poids qui respectent l'équation de calage.

6. ESTIMATEURS POUR DOMAINES AVEC UNE COMPOSANTE SYNTHÉTIQUE

On utilise des poids restreints à cause des propriétés de l'estimateur par calage pour de petites tailles d'échantillon. Pour de grandes tailles d'échantillon, on sait qu'on a normalement $w^{\text{cal}} - A^s c_s$ qui converge en probabilité vers

0, donc des poids qui ne sont pas problématiques. Un statisticien qui est confronté au problème de poids extrêmes doit donc vraisemblablement faire face à un autre problème lié aux petites tailles d'échantillon, à savoir l'estimation pour de petits domaines. On présentera dans cette section, un estimateur dont les propriétés asymptotiques sont celles de l'estimateur par calage, mais qui utilise des poids restreints et tire profit d'un estimateur synthétique.

Soit $\tilde{X} = X\tilde{\beta}_s$ une estimation synthétique pour X , on a

$$\tilde{X}^s w_s = (X^s \tilde{\beta}_s) w_s$$

$$= \tilde{\beta}_s^s X^s w_s$$

$$= \tilde{\beta}_s^s X^s c$$

$$= (X\tilde{\beta}_s)^s c$$

$$= \tilde{Y}^s c$$

(13)

avec égalité pour la troisième expression de droite si les poids satisfont l'équation de calage $X^s w_s = X^s c$. Les poids w^{cal} données par (1) minimisent $\|X^s w_s - X^s c\|_T^2$. On peut donc estimer $\tilde{Y}^s c$ par

$$\hat{\tau} = (X^s - \tilde{X}^s) w_s^{\text{res}} + \tilde{Y}^s c. \quad (14)$$

Il y aura égalité entre cet estimateur et l'estimateur $\tilde{Y}^s w^{\text{cal}}$ dès que l'échantillon sera suffisamment grand pour que l'équation de calage soit satisfaite et pour que w^{cal} soit dans la région de restriction, c'est-à-dire dès que $w^{\text{res}} = w^{\text{cal}}$.

Les propriétés asymptotiques de ces deux estimateurs sont donc les mêmes sous certaines conditions discutées dans la section précédente. L'avantage de l'estimateur $\hat{\tau}$ est qu'il donne une estimation synthétique lorsque des colonnes de X^s et \tilde{X}^s sont nulles.

7. DONNÉES ABERRANTES

On pourrait traiter les données aberrantes d'une façon similaire aux poids extrêmes. La stratégie est la suivante: on se donne une région de restriction pour $\tilde{X}^s w^{\text{cal}}$, on montre que pour n suffisamment grand $\tilde{X}^s w^{\text{cal}}$ est à l'intérieur de cette région de restriction, et on se donne un estimateur plus «raisonnable» pour remplacer $\tilde{X}^s w^{\text{cal}}$ dans les cas où $\tilde{X}^s w^{\text{cal}}$ est à l'extérieur de la région de restriction. Dans le cas d'un échantillon stratifié on aurait normalement une région de restriction par strate.

On a montré à la section 2 que sous certaines conditions sur le cadre asymptotique, $w^{\text{cal}} - A^s c_s = O_p(n^{-1/2} N_Y)$. On a donc $\tilde{X}^s w^{\text{cal}} - \tilde{X}^s A^s c_s = O_p(n^{-1/2} N_Y)$, et si on suppose

$$X^s A^s c_s - A^s c_s = O_p(n^{-1/2} N_Y), \quad (15)$$

alors $\tilde{X}^s w^{\text{cal}} - A^s c_s = O_p(n^{-1/2} N_Y)$. Un expert (ou un groupe d'experts) peut déterminer à partir d'informations indépendantes des données d'enquête, qu'il ne serait pas

où $Y^* = X\beta(\alpha)$, Π est la matrice des probabilités d'inclusion de second ordre, et $\text{diag}(c)$ la matrice diagonale formée à partir du vecteur c .

5. MÉTHODES D'ESTIMATION AVEC POIDS RESTREINTS

Afin d'éviter d'obtenir des poids ayant des valeurs extrêmes, on peut vouloir forcer le vecteur de poids à être à l'intérieur d'une région déterminée. On supposera que A_s^c est un point de cette région. Par exemple, si $w_s^{(L)} < A_s^c < w_s^{(H)}$, on pourrait vouloir restreindre les poids à la région $R_w^* = \{w_s : w_s^{(L)} \leq w_s \leq w_s^{(H)}\}$. On supposera que

$$\lim_{n \rightarrow \infty} w_s^{(L)} - A_s^c < 0 \text{ et } \lim_{n \rightarrow \infty} w_s^{(H)} - A_s^c > 0.$$

L'approche mentionnée à la section 3 consiste à choisir une mesure de distance entre les poids calés et les poids de Horvitz-Thompson qui donnera des poids satisfaisant l'équation de calage qui sont dans la région de restriction, ceci pourvu que de tels poids existent. L'approche qui sera étudiée dans cette section consiste à tempérer l'exigence de satisfaisance l'équation de calage lorsque le vecteur des poids de calage w^{cal} est à l'extérieur de la région de restriction. Différentes façons de tempérer cette exigence conduisent à différentes méthodes de pondération.

Lorsque w^{cal} est à l'extérieur de la région de restriction, on pourrait par exemple, chercher les points de la courbe $w_s(\alpha)$ paramétrisée par $\alpha \geq 0$ qui sont sur la frontière de cette région. Ces points ont la propriété d'être une solution du problème de minimisation décrit dans la section 4 pour les valeurs de α correspondantes, donc à travers la matrice T , on peut pondérer l'importance de chaque équation de calage. Avec l'exemple de la région de restriction donné plus haut, si

$$w^{\text{cal}} = \lim_{\alpha \rightarrow \infty} w_s(\alpha)$$

est à l'intérieur de cette région, alors on peut prendre comme vecteur de poids restreints $w^{\text{res}_1} = w^{\text{cal}}$, sinon on peut prendre $w^{\text{res}_1} = w_s(\alpha)$ pour $\alpha < \infty$ tel que $w_s(\alpha)$ est à la frontière de la région de restriction. Si le cadre asymptotique est tel que les conditions (2) sont respectées avec $\gamma < 3/2$ alors pour n suffisamment grand, la probabilité que w^{cal} soit à l'intérieur de la région de restriction est égale à 1. En effet, on a $w^{\text{cal}} - A_s^c$ converge en probabilité vers 0. Les propriétés asymptotiques de l'estimateur qui utilise les poids restreints, w^{res_1} , sont donc les mêmes que celles de l'estimateur par calage. Il est important de noter que puis-que $|w_s - a_s^c|$ n'est pas nécessairement une fonction monotone de α , il est possible que $w_s(\alpha)$ soit à la frontière de la région de restriction pour plusieurs valeurs de α , même si la région de restriction est convexe. Il n'est pas

nécessairement simple de trouver toutes ces valeurs, et il faut décider laquelle utiliser.

Une autre option pour restreindre les poids consisterait à se donner comme région de restriction, les poids w_s qui satisfont $D_s(w_s) \leq l$ pour une borne $l > 0$. On prend ensuite comme vecteur de poids restreints $w^{\text{res}_2} = w^{\text{cal}}$, si w^{cal} est dans la région de restriction, sinon on cherche $\alpha > 0$ tel que $D_s(w_s(\alpha)) = l$. Cette valeur de α est unique et peut être trouvée de façon itérative. On calcule ensuite les poids

$w^{\text{res}_2} = w_s(\alpha)$ qui correspondent à cette valeur de α à l'aide de l'équation (12). Si le cadre asymptotique est tel que les conditions (2) sont respectées avec $\gamma < 1$, et si l ne varie pas avec n , alors pour n suffisamment grand, la probabilité que w^{cal} soit à l'intérieur de la région de restriction est égale à 1. En effet, on a $D_s(w^{\text{cal}})$ converge en probabilité vers 0. Les propriétés asymptotiques de l'estimateur qui utilise les poids restreints, w^{res_2} , sont alors les mêmes que celles de l'estimateur par calage. Malheureusement, lorsqu'on estime un total il faut s'attendre à avoir $\gamma = 1$. Pour pallier cet inconvénient, on peut utiliser l/\sqrt{n} comme borne, plutôt que l . On peut justifier cette borne par le fait que la longueur de la diagonale principale d'un hypercube de \mathbb{R}^n est égale au diamètre de la boule qui circonscrit cet hypercube, par contre, le diamètre de la boule inscrite dans ce même hypercube est plus petit par un facteur de \sqrt{n} . Il reste que le statisticien peut être inconfortable avec l'utilisation d'un cadre asymptotique où la borne croît avec la taille de l'échantillon. Il y a aussi le fait que cette approche ne permette pas de limiter individuellement les poids des observations. Seule la distance entre le vecteur des poids restreints et le vecteur des poids de Horvitz-Thompson est contrôlée.

Avec les méthodes décrites plus haut, on cherche les points de la courbe $w_s(\alpha)$ qui sont sur la frontière de la région de restriction. La valeur de α pour laquelle $w_s(\alpha)$ est à la frontière de la région de restriction doit souvent être trouvée de façon itérative. On pourrait plus simplement remplacer la courbe $w_s(\alpha)$ par le segment de droite reliant A_s^c à w^{cal} . Pour la région de restriction R_w , on aurait comme vecteur de poids restreints $w^{\text{res}_3} = w^{\text{cal}}$, si w^{cal} est dans la région de restriction, et sinon w^{res_3} serait égal au point où le segment de droite traverse la frontière de la région de restriction, c'est-à-dire

$$w^{\text{res}_3} = A_s^c + \xi(w^{\text{cal}} - A_s^c),$$

où

$$\xi = \min_k \{ \max [(w_s^{(L)} - A_s^c) / (w_s^{\text{cal}} - A_s^c), (w_s^{(H)} - A_s^c) / (w_s^{\text{cal}} - A_s^c)] \},$$

la division des vecteurs se faisant élément par élément, le maximum des deux vecteurs étant pris élément par élément, et \min donnant l'élément minimum. On pourrait aussi considérer le vecteur de poids de la région de restriction, w^{res_4} ,

où

$$V = \begin{pmatrix} 0 & \alpha I \\ U_s & 0 \end{pmatrix}$$

et $\alpha \geq 0$. On minimise alors

$$\|w_s - A_s c\|_2^2 + \alpha \|X_s' w_s - X_s' c\|_2^2 =$$

$$D_s(w_s) + \alpha \|X_s' w_s - X_s' c\|_2^2.$$

Un problème de minimisation semblable est rencontré lors d'une régression avec coefficients de coûts (ridge regression). Pour $\alpha = 0$ la solution est donnée par les poids de Horvitz-Thompson $w_s = A_s c$. Tandis que pour $\alpha > 0$, on cherche $w_s(\alpha)$ qui minimise $\|K(w_s - A_s c) - b\|_2^2$, où $K = (U_n, X_s')$, $b = (0^{1 \times n}, (X_s' c - X_s' A_s c)')$ et $0^{1 \times n} \in \mathbb{R}^n$ est un vecteur-ligne de 0. Ben-Israel et Greville (1980) donne

$$w_s(\alpha) - A_s c = (K'VK)^{-1}K'Vb. \quad (10)$$

Donc on obtient en substituant les valeurs de K , V , et b

$$w_s(\alpha) = A_s c + \alpha(U_s' -$$

$$+ \alpha X_s' T X_s')^{-1} X_s' T(X_s' c - X_s' A_s c). \quad (11)$$

On montre facilement que

$$\alpha(U_s' + \alpha X_s' T X_s')^{-1} X_s' T$$

le quotient

$$= U_s^{-1} X_s(\alpha^{-1} I - 1 + X_s' U_s^{-1} X_s)^{-1},$$

$$Y_s' A_s c_s +$$

$$[(X_s' A_s 1^{n \times 1}) / (X_s' A_s 1^{n \times 1})](X_s' c - X_s' A_s c_s),$$

où $1^{n \times b} \in \mathbb{R}^{n \times b}$ est une matrice de 1.

Ben-Israel et Greville (1980, 111, exercice 15) montrent que $D_s(w_s(\alpha))$ est une fonction monotone croissante de α .

Il faut cependant noter que pour une unité $k \in s$, $|w_k(\alpha) - a_k c_k|$ n'est pas nécessairement une fonction monotone de α . Lorsque α augmente, le vecteur de poids $w_s(\alpha)$ s'éloigne du vecteur des poids de Horvitz-Thompson, mais ce n'est pas nécessairement le cas pour

chaque coordonnée de ce vecteur.

Dans cet article on utilisera le calage mitigé pour restreindre les poids, donc lorsque la taille d'échantillon est relativement petite. Il est cependant facile de montrer que $\beta_s(\alpha) - \beta(\alpha) \rightarrow 0$ en probabilité, avec

$$\beta(\alpha) = (X'U^{-1}X + \alpha^{-1}T^{-1})^{-1}X'U^{-1}Y,$$

on a $w_s(\alpha)$ est un estimateur asymptotiquement non biaisé dont la variance asymptotique est

$$(X - X^*)' \text{diag}(c)(A \Pi A - 1^{N \times N}) \text{diag}(c)(X - X^*),$$

$$\lim_{\alpha \rightarrow \infty} w_s(\alpha) = w_{\text{cal}}.$$

avec $F = I^{1/2} X_s' U_s^{-1/2}$, pour montrer que

À partir de l'équation (12), on peut également utiliser Ben-Israel et Greville (1980), et le fait que $F' = F'(FF')^{-1}$

régression avec coefficients de coûts.

parallelisable semblable peut être fait entre le calage mitigé et la régression par calage et la méthode d'estimation par régression généralisée telle que décrite dans Särndal, Swensson et Wretman (1992), conduisent aux mêmes estimateurs, un qu'on obtient lors d'une régression avec coefficients de coûts (ridge regression). Tout comme la méthode d'esti-

$$\beta_s(\alpha) = (X_s' U_s^{-1} X_s + \alpha^{-1} T^{-1})^{-1} X_s' U_s^{-1} Y_s.$$

L'estimateur $X_s' w_s(\alpha)$ prend donc la forme

$$+ X_s' U_s^{-1} X_s)^{-1} (X_s' c - X_s' A_s c_s). \quad (12)$$

$$w_s(\alpha) = A_s c_s + U_s^{-1} X_s(\alpha^{-1} T^{-1}$$

d'où

(9)

$$\sum_{j \in T} \left(N_{(L)}^f - \sum_{i \in S} N_{(H)}^f \right) \leq \sum_{i \in S} \left(N_{(L)}^f - \sum_{j \in T} N_{(H)}^f \right)$$

$$\sum_{i \in S} \left(N_{(L)}^f - \sum_{j \in T} N_{(H)}^f \right) + \sum_{j \in T} \left(N_{(L)}^f - \sum_{i \in S} N_{(H)}^f \right) \leq \sum_{i \in S} \left(N_{(L)}^f - \sum_{j \in T} N_{(H)}^f \right) + \sum_{j \in T} \left(N_{(L)}^f - \sum_{i \in S} N_{(H)}^f \right)$$

$$\sum_{i \in S} \left(N_{(L)}^f - \sum_{j \in T} N_{(H)}^f \right) + \sum_{j \in T} \left(N_{(L)}^f - \sum_{i \in S} N_{(H)}^f \right) \leq \sum_{i \in S} \left(N_{(L)}^f - \sum_{j \in T} N_{(H)}^f \right) + \sum_{j \in T} \left(N_{(L)}^f - \sum_{i \in S} N_{(H)}^f \right)$$

pour tout $S \subseteq \{1, 2, \dots, R\}$, $T \subseteq \{1, 2, \dots, C\}$.

On peut réduire le nombre d'inéquations à vérifier. Par exemple, plutôt que de vérifier

$$\sum_{j \in T} \left(N_{(L)}^f - \sum_{i \in S} N_{(H)}^f \right) \leq \sum_{i \in S} \left(N_{(L)}^f - \sum_{j \in T} N_{(H)}^f \right)$$

pour tout $S \subseteq \{1, 2, \dots, R\}$, et $T \subseteq \{1, 2, \dots, C\}$, on montre facilement que il est équivalent de vérifier que

$$\sum_{j \in T} \left(N_{(L)}^f - \sum_{i \in S} N_{(H)}^f \right) \leq \sum_{i \in S} \left(N_{(L)}^f - \sum_{j \in T} N_{(H)}^f \right)$$

pour tout $T \subseteq \{1, 2, \dots, C\}$.

4. CALAGE MITIGÉ

si et seulement si

$$N_{(L)}^f \leq \sum_{j=1}^R \sum_{i=1}^C N_{(H)}^f, \quad i=1, \dots, R, \quad j=1, \dots, C,$$

$$N_{(L)}^f \leq \sum_{j=1}^R N_{(H)}^f, \quad i=1, \dots, R, \quad j=1, \dots, C,$$

$$N_{(L)}^f \leq \sum_{j=1}^R N_{(H)}^f, \quad i=1, \dots, R, \quad j=1, \dots, C,$$

$$N_{(L)}^f \leq N_{(H)}^f, \quad i=1, \dots, R, \quad j=1, \dots, C,$$

Ces conditions sont dans une certaine mesure redondantes. Par exemple, si les inégalités (7) sont satisfaites pour $A^{\text{sub}} = V_1$, alors elles sont nécessairement satisfaites pour toute matrice V_2 obtenue de V_1 par une permutation de lignes. Un autre exemple est fourni par la pondération d'observations dans un tableau de contingence. Soient $N_{ij}^f = n_{ij}^f$ ($i = 1, 2, \dots, R; j = 1, 2, \dots, C$), où n_{ij}^f est le nombre d'observations dans la cellule (i, j) du tableau de contingence et w_{ij}^f est le poids de chacune de ces observations, on veut savoir s'il existe des poids w_{ij}^f tels que N_{ij}^f satisfait certaines contraintes. Par exemple, motivé par le problème de la convergence de la procédure du quotient réitéré (raking ratio), Bacharach (1965) donne des conditions nécessaires et suffisantes pour l'existence de poids w_{ij}^f tels que $N_{ij}^f \geq 0$, $\sum_{j=1}^C N_{ij}^f = N_i^f$ ($i = 1, \dots, R$), où les valeurs de N_i^f et N_{ij}^f sont données. Le résultat suivant qui est démontré à l'annexe B, est plus général. Pour des constantes arbitraires $N_{(L)}^f, N_{(H)}^f, N_{(L)}^f, N_{(H)}^f$, et $N_{(H)}^f$, il existe des N_{ij}^f tels que

$$N_{(L)}^f \leq N_{(H)}^f, \quad i=1, \dots, R, \quad j=1, \dots, C,$$

$$N_{(L)}^f \leq N_{(H)}^f, \quad j=1, \dots, C,$$

$$N_{(L)}^f \leq N_{(H)}^f, \quad i=1, \dots, R,$$

$$N_{(L)}^f \leq N_{(H)}^f$$

On peut ne pas être satisfait avec l'approche en deux temps de l'estimation par calage, où on cherche d'abord les vecteurs de poids qui satisfont l'équation de Horvitz-Thompson. Pour de petits échantillons, cette méthode peut conduire à des poids que le statisticien considère trop éloignés des poids de Horvitz-Thompson. On pourrait préférer doser l'importance qu'on accorde à l'équation de calage par rapport à la norme de $w_s - A_s^s$. Ainsi, on peut désirer trouver un vecteur de poids w_s qui minimise

$$\left\| \begin{pmatrix} w_s - A_s^s \\ X_s' w_s - X_s' c \end{pmatrix} \right\|_V^2$$

Corollaire. Soient $M \in \mathbb{R}^{m \times n}$ et $L, h \in \mathbb{R}^m$, $\exists w \in \mathbb{R}^n$ tel que $L < Mw < h$ si et seulement si premièrement $L < h$ et deuxièmement $\lambda \in N(M') \rightarrow -L' \lambda_- < h' \lambda_+$, où $\lambda_+ = \max(\lambda, 0)$ et $\lambda_- = \min(\lambda, 0)$ les extrema étant pris élément par élément.

Le corollaire est obtenu en posant

$$M' = \begin{pmatrix} M \\ -M \end{pmatrix}, L' = \begin{pmatrix} L \\ -h \end{pmatrix} \text{ et } \lambda = \begin{pmatrix} \lambda_+ \\ \lambda_- \end{pmatrix}$$

dans le théorème.

On note p la dimension de $N(M')$. Si p est égal à zéro, alors $\lambda \in N(M')$ implique $\lambda = 0$, et la condition du théorème (ou la condition du corollaire) est évidemment satisfaite. Si p est égal à un, alors $\lambda \in N(M')$ implique que λ est un multiple d'un vecteur z , et il est suffisant de vérifier la condition pour $\lambda = z$ et $\lambda = -z$. En utilisant la propriété $(-\lambda)_+ = -(-\lambda)_-$, le problème posé au début de la section peut maintenant être résolu si X_s est un vecteur. Le corollaire avec

$$M' = \begin{pmatrix} I_n \\ X_s' \end{pmatrix}, L' = \begin{pmatrix} I_{(T)} \\ I_{(H)} \end{pmatrix}, h = \begin{pmatrix} I_{(T)} \\ I_{(H)} \end{pmatrix}$$

et le fait que

$$\begin{pmatrix} I \\ -X_s' \end{pmatrix} = z$$

engendre $N(M')$, donne les conditions nécessaires et suffisantes

$$(5) \quad \begin{aligned} I_{(H)} w &\leq w_{(H)} \\ I_{(T)} I &\leq I_{(H)} I \\ I_{(H)} I_{(H)}' w_{(H)} + (X_s')_{(H)}' w_{(H)} &\leq I_{(T)} I \end{aligned}$$

La troisième inégalité dans (5) affirme que le total pondéré de la variable auxiliaire ne doit pas être supérieur à $I_{(H)}$, lorsque le plus petit poids possible $w_{(T)}$, est donné aux unités où la variable auxiliaire prend une valeur positive, et lorsque le plus grand poids possible $w_{(H)}$, est donné aux unités où la variable auxiliaire prend une valeur négative. La quatrième inégalité dans (5) affirme que le total pondéré de la variable auxiliaire ne doit pas être inférieur à $I_{(T)}$, lorsque le plus grand poids possible $w_{(H)}$, est donné aux unités où la variable auxiliaire prend une valeur positive, et lorsque le plus petit poids possible $w_{(T)}$, est donné aux unités où la variable auxiliaire prend une valeur négative. Même lorsque $p > 1$, il est suffisant de vérifier la condition du corollaire pour un nombre fini de valeurs de λ . Soit $V \in \mathbb{R}^{m \times p}$, $2 \leq p \leq m$ une matrice dont les colonnes forment une base pour $N(M')$. Il est toujours possible de construire V de sorte que p de ses lignes, v_1, v_2, \dots, v_m , soient les vecteurs unités de \mathbb{R}^p , et on supposera que V est

de cette forme. On démontre dans l'annexe A qu'il est suffisant de vérifier la condition du corollaire pour les vecteurs $\lambda = V\phi$ et $\lambda = -V\phi$, où $\phi = (\phi_1, \dots, \phi_p)'$ est un vecteur non nul et satisfait $v_i' \phi = 0$ pour un sous-ensemble de $(p-1)$ vecteurs v_i linéairement indépendants. On doit donc vérifier la condition pour au plus $\binom{p-1}{m}$ vecteurs ϕ , donc au plus $2 \binom{p-1}{m}$ valeurs de λ .

En utilisant le corollaire avec

$$M' = \begin{pmatrix} I_n \\ X_s' \end{pmatrix}, L' = \begin{pmatrix} I_{(T)} \\ I_{(H)} \end{pmatrix}, h = \begin{pmatrix} I_{(T)} \\ I_{(H)} \end{pmatrix},$$

et en notant que les colonnes de

$$A = \begin{pmatrix} -X_s' \\ I_p \end{pmatrix}$$

forment une base pour $N(M')$, on obtient les conditions nécessaires et suffisantes suivantes pour l'existence d'une solution au problème mentionné au début de cette section lorsque $X_s \in \mathbb{R}^{n \times p}$ avec $p > 1$. On doit avoir $w_{(T)} I_{(T)} \leq I_{(H)} I_{(H)}'$, et pour chaque sous-ensemble de $(p-1)$ lignes

de

$$A' = \begin{pmatrix} -X_s' \\ I_p \end{pmatrix}$$

linéairement indépendantes,

$$(6) \quad \begin{aligned} I_{(H)} I_{(H)}' w_{(H)} + (X_s')_{(H)}' w_{(H)} &\leq I_{(T)} I \\ I_{(T)} I &\leq I_{(H)} I \\ I_{(H)} I_{(H)}' \phi - \phi_{(H)}' w_{(H)} &\leq I_{(T)} I \end{aligned}$$

pour un vecteur non nul $\phi \in \mathbb{R}^p$ de direction perpendiculaire à chaque ligne du sous-ensemble. La deuxième inéquation de (6) est obtenue de la première en changeant le signe de ϕ .

Si $V^{\text{sup}} \in \mathbb{R}^{p \times p}$ est une matrice non singulière dont les lignes sont des lignes de V , alors chaque colonne de V^{sup} est un vecteur perpendiculaire à $(p-1)$ lignes de V linéairement indépendantes. D'où le résultat suivant:

Il existe un vecteur de poids w_s tels que $w_{(T)} I_{(T)} \leq w_s \leq w_{(H)} I_{(H)}'$ si et seulement si $w_{(H)} I_{(H)}' \leq w_{(T)} I_{(T)}$ et

$$(7) \quad \begin{aligned} (X_s')_{(H)}' w_{(H)} - (X_s')_{(T)}' w_{(T)} &\leq - (X_s')_{(H)}' w_{(H)} \\ (X_s')_{(H)}' w_{(H)} + (X_s')_{(T)}' w_{(T)} &\leq (X_s')_{(H)}' w_{(H)} \end{aligned}$$

pour toutes les matrices non singulières $V^{\text{sup}} \in \mathbb{R}^{p \times p}$ dont les lignes sont des lignes de

auxiliaire $X \in \mathbb{R}^{N \times p}$, $A \in \mathbb{R}^{N \times N}$ la matrice diagonale des poids de sondage, des matrices diagonales positives données $U \in \mathbb{R}^{n \times n}$ et $T \in \mathbb{R}^{p \times p}$, on cherche parmi les vecteurs de poids $w_s \in \mathbb{R}^n$ qui minimisent $\|X_s^T w_s - X^T\|_T^2$, celui qui minimise $D_s(w_s) = \|w_s - A_s c\|_{U_s}^2$. Cette formulation du problème qui ne présume pas de l'existence de poids satisfaisant l'équation de calage, $X_s^T w_s = X^T c$, est donnée dans Théberge (1999). La solution recherchée constitue le vecteur des poids calés w^{cal} . On a

$$w^{\text{cal}} = A_s c + U_s^{-1} X_s^T T_{1/2} (T_{1/2} X_s^T U_s^{-1} X^T T_{1/2})^{\dagger}$$

$$T_{1/2} (X^T c - X_s^T A_s c), \quad (1)$$

où F^{\dagger} dénote l'inverse de Moore-Penrose de la matrice F .

Afin de mieux étudier les propriétés asymptotiques d'estimateurs par calage avec poids restreints, on examinera maintenant le comportement de w^{cal} lorsque $n \rightarrow \infty$. On suppose l'existence d'un cadre asymptotique où la taille de la population et la taille de l'échantillon tendent vers l'infini, voir par exemple Isaki et Fuller (1982), et pour lequel on a

$$X^T c = O^p(N^{\gamma}) \quad (\gamma \geq 0)$$

$$X^T c - X_s^T A_s c = O^p(n^{-1/2} N^{\gamma}) \quad (2)$$

$$T_{1/2} X_s^T U_s^{-1} X_s^T T_{1/2} = O^p(n).$$

Il s'ensuit que $(T_{1/2} X_s^T U_s^{-1} X_s^T T_{1/2})^{\dagger} = O^p(n^{-1})$, puisqu'une des propriétés de l'inverse de Moore-Penrose d'une matrice F est $F^{\dagger} F F^{\dagger} = F^{\dagger}$. Habituellement, on peut s'attendre à avoir $\gamma = 1$ lorsque chaque élément du vecteur c vaut 1 (estimation d'un total), et $\gamma = 0$ lorsque chaque élément de c vaut $1/N$ (estimation d'une moyenne). Sous les conditions (2) on a donc,

$$w^{\text{cal}} - A_s c = U_s^{-1} X_s^T T_{1/2} (T_{1/2} X_s^T U_s^{-1} X_s^T T_{1/2})^{\dagger}$$

$$T_{1/2} (X^T c - X_s^T A_s c)$$

$$= O^p(n^{-1}) O^p(n^{-1/2} N^{\gamma})$$

$$= O^p(n^{-3/2} N^{\gamma}). \quad (3)$$

Donc $w^{\text{cal}} - A_s c$ converge en probabilité vers 0, si

$$\lim_{n, N \rightarrow \infty} n^{-3/2} N^{\gamma} = 0.$$

Pour un cadre asymptotique tel celui de Brewer (1979) où la fraction de sondage n/N est constante, ou tout cadre pour lequel la fraction de sondage converge vers un nombre positif, cette condition est vérifiée si $\gamma < 3/2$.

Si on écrit $w^{\text{cal}} = A_s c + U_s^{-1} X_s^T T_{1/2} H_s^{\dagger} T_{1/2} (X^T c - X_s^T A_s c)$, où $H_s = T_{1/2} X_s^T U_s^{-1} X_s^T T_{1/2}$, on a

$$a \quad 1/\lambda \leq 0.$$

Théorème: Soient $M \in \mathbb{R}^{m \times n}$ et $l \in \mathbb{R}^m$, $\exists w \in \mathbb{R}^n$ tel que

$Mw \leq l$ si et seulement si pour tout $\lambda \geq 0$ dans $N(M')$, on

a $l^T \lambda \leq 0$.
L'ensemble des vecteurs a tels que $M'a = 0$.
théorème fait appel au noyau de M' , $N(M')$, défini comme

dimension finie, quoique la preuve donnée par Fan tiennne

également pour une matrice de dimension infinie. Le

Une première étape est fournie par le théorème suivant

de Fan (1956). Il est énoncé ici pour une matrice M de

pour l'existence de poids restreints aux intervalles

posant $t^{(T)} = t^{(H)} = X^T c$, on obtiendrait des conditions

où $w^{(T)}$, $w^{(H)}$ et $t^{(H)}$ sont donnés. En particulier, en

d'un vecteur w_s tel que $w_s^{(T)} \leq w_s^{(H)}$ et $t^{(T)} \leq t^{(H)}$,

C'est-à-dire qu'on cherche des conditions pour l'existence

totaux pour les variables auxiliaires sont également bornées.

rietur de bornes données, et tel que les estimations des

santes pour l'existence d'un vecteur de poids w_s à l'inté-

section est de trouver des conditions nécessaires et suffi-

qui satisfait l'équation de calage. Le but principal de cette

fonctionnera que s'il existe des poids dans ces intervalles

en satisfaisant l'équation de calage. Cette approche ne

calés, afin de restreindre les poids à certains intervalles tout

distance entre les poids de Horvitz-Thompson et les poids

autres qu'une somme pondérée de carrés pour mesurer la

(1992) ont proposé d'utiliser diverses mesures de distance

négatifs ou exceptionnellement grands. Deville et Samdal

solution existe, il est possible que les poids calés soient

une solution si et seulement si $(X_s^T X_s^T)' X^T c = X^T c$. Si une

calage, on obtient que l'équation de calage $X_s^T w_s = X^T c$ a

calage. En appliquant Graybill (1983, 113) au problème de

est possible qu'il n'y ait pas de solution à l'équation de

Même lorsqu'il n'y a pas de restrictions sur les poids, il

3. SOLUTIONS À L'ÉQUATION DE CALAGE

ET POIDS RESTREINTS

$\|w^{\text{cal}} - A_s c\|_{U_s}^2$ ne converge pas vers 0.

converge en probabilité vers 0, mais où $D_s(w^{\text{cal}}) =$

notamment pour l'estimation d'un total, où $w^{\text{cal}} - A_s c$

converge en probabilité vers 0, si $\gamma < 1$. Il y a donc des cas,

sondage converge vers un nombre positif, on a $D_s(w^{\text{cal}})$

Toujours pour un cadre asymptotique où la fraction de

$D_s(w^{\text{cal}}) = (X^T c - X_s^T A_s c)' T_{1/2} H_s^{\dagger} H_s^{\dagger} T_{1/2} (X^T c - X_s^T A_s c)$

$T_{1/2} (X^T c - X_s^T A_s c)$

$D_s(w^{\text{cal}}) = (X^T c - X_s^T A_s c)' T_{1/2} H_s^{\dagger} H_s^{\dagger} T_{1/2} (X^T c - X_s^T A_s c)$

$(X^T c - X_s^T A_s c)$

$(X^T c - X_s^T A_s c)$

$(X^T c - X_s^T A_s c)$

$(X^T c - X_s^T A_s c)$

Calage et poids restreints

ALAIN THÉBERGE¹

RÉSUMÉ

Pour mieux comprendre l'impact de l'imposition d'une région de restriction sur les poids de calage, on examine le comportement asymptotique de ceux-ci. On donne des conditions nécessaires et suffisantes pour l'existence d'une solution à l'équation de calage avec des poids à l'intérieur d'intervalles donnés. Une formulation plus générale du problème de calage permet de faire un compromis entre le besoin de satisfaire l'équation de calage, et le désir d'obtenir des poids qui sont près des poids de Horvitz-Thompson. Si on relâche les exigences vis-à-vis l'équation de calage, alors diverses méthodes d'estimation avec poids restreints peuvent être utilisées. Les estimateurs présents ont habituellement les mêmes propriétés asymptotiques que l'estimateur par calage sans restrictions sur les poids, et certains ont des poids qui peuvent être calculés explicitement, sans procédure itérative. On montre comment ces estimateurs peuvent être adaptés pour tirer parti d'un estimateur synthétique. Une stratégie semblable à celle utilisée pour restreindre les poids est appliquée aux données aberrantes.

MOTS CLÉS : Petits domaines; inverse de Moore-Penrose; solutions d'inéquations; propriétés asymptotiques; données aberrantes.

1. INTRODUCTION

L'estimateur par calage possède de bonnes propriétés asymptotiques. Mais pour de petites tailles d'échantillon, ou si le calage se fait au niveau de domaines dont certains ont peu d'observations, les poids de cet estimateur peuvent prendre des valeurs extrêmes. Une façon de pallier ce problème consiste à utiliser la méthode de calage avec des mesures de distance qui font que les poids des observations sont restreints à certains intervalles autour des poids de Sarnadal (1992). D'autres méthodes qui visent à obtenir des estimations robustes qui respectent l'équation de calage sont données dans Duchesne (1999). Cet article contient une bonne bibliographie sur les estimateurs robustes. On n'est toutefois pas assuré de l'existence d'une solution à l'équation de calage avec des poids restreints. Même si de tels poids existent, il se peut que le statisticien préfère résoudre le problème de poids extrêmes en relâchant quelque peu ses exigences vis-à-vis l'équation de calage, plutôt que de distance plus «contraignante». On présentera ici, une formulation du problème de calage qui offre plus de flexibilité au statisticien. Il s'agit en fait d'un problème de minimisation du même ordre que celui rencontré lors d'une régression avec coefficients de coûts («ridge regression»). Bardley et Chambers (1984) ont rencontré le même problème de minimisation dans leur recherche d'estimateurs basés sur des modèles. Cette formulation du problème de calage peut être utilisée pour restreindre les poids sans utiliser de mesure de distance spéciale entre les poids calés et les poids de Horvitz-Thompson. Rao et Singh (1997) ont eux combiné cette approche avec des méthodes itératives

utilisant des mesures de distance. D'autres façons de restreindre les poids seront aussi examinées. Dans la section suivante on présente la méthode de calage en l'absence de bornes sur les valeurs des poids. Le problème de calage qui est posé ne présuppose pas de l'existence d'une solution à l'équation de calage. Les propriétés asymptotiques des poids calés sont discutées. Ces propriétés sont d'intérêt pour le comportement asymptotique des estimateurs dont les poids sont restreints. La section 3 donne des conditions nécessaires et suffisantes pour l'existence de poids restreints qui satisfont l'équation de calage. À la section 4, on discute d'une façon de poser le problème d'estimation en dosant l'importance qu'on accorde à l'équation de calage. La section 5 donne différentes façons de restreindre les poids qui ne reposent pas sur l'utilisation d'une distance particulière. On propose à la section 6, un estimateur avec poids restreints qui est utile pour de petits domaines. Finalement, la section 7 aborde les données aberrantes en développant une méthode semblable à celle utilisée pour traiter les poids extrêmes.

2. CALAGE

Soient $Y \in \mathbb{R}^{N \times d}$ une matrice de d variables d'intérêt pour une population de taille N , et $c \in \mathbb{R}^N$ un vecteur de constantes connues, on tire un échantillon s de taille n et on notera à l'aide de l'indice s les sous-vecteurs ou les sous-matrices qui correspondent à l'échantillon. Il s'agit d'estimer $Y'c$ par $Y'_s w_s$, où $w_s \in \mathbb{R}^n$ est un vecteur de poids pour les unités échantillonnées. Pour un vecteur v et une matrice diagonale positive F de même dimension, on définit $\|v\|_F^2 = v' F v$. Pour une matrice d'information

¹ Alain Théberge, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa (Ontario) K1A 0T6 Canada.

Thompson et Frank: Estimation fondée sur un modèle et comportant des plans d'échantillonnage à dépistage 112

WASSERMAN, S., et FAUST, K. (1994). *Social Network Analysis: Methods and Applications*. New York: Cambridge University Press.

WATTERS, J.K., et BIERNACKI, P. (1989). Targeted sampling: Options for the study of hidden populations. *Social Problems*, 36, 416-430.

WELLMAN, B., FRANK, O., ESPINOZA, V., LUNDQVIST, S., et WILSON, C. (1991). Integrating individual, relational and structural analysis. *Social Networks*, 13, 223-249.

WEI, L.J., SMYTHE, R.T., LIN, D.Y., et PARK, T.S. (1990). Statistical inference with data-dependent treatment allocation rules. *Journal of the American Statistical Association*, 85, 156-162.

- JANSSON, I. (1997). On statistical modeling of social networks. Thèse de doctorat. Stockholm University.
- KALTON, G., et ANDERSON, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society A*, 149, 65-82.
- KARLBERG, M. (1997). Triad count estimation and transitivity testing in graphs and digraphs. Thèse de doctorat. Stockholm University.
- KLOVD AHL, A.S. (1989). Urban social networks: Some methodological problems and possibilities. Dans *The Small World*, (Ed. M. Kochen). Norwood, NJ: Ablex Publishing, 176-210.
- LEJEUNE, M., et FAULKENBERG, G.D. (1982). A simple predictive density function. *Journal of the American Statistical Association*, 77, 654-657.
- LEVY, P.S. (1977). Optimum allocation in stratified random network sampling for estimating the prevalence of attributes in rare populations. *Journal of the American Statistical Association*, 72, 758-763.
- LEVY, P.S., et LEMESHOW, S. (1991). *Sampling of Populations: Methods and Applications*. New York: Wiley.
- LINDSAY, B.G., et LI, B. (1997). On second-order optimality of the observed Fisher information. *Annals of Statistics*, 25, 2172-2199.
- MORGAN, D.L., et RYTINA, S. (1977). Comment on "Network sampling: some first steps" by Mark Granovetter. *American Journal of Sociology*, 83, 722-727.
- NEAIGUS, A., FRIEDMAN, S.R., GOLDSTEIN, M.F., ILDEFONSO, G., CURTIS, R., et JOSE, B. (1995). Using dyadic data for a network analysis of HIV infection and risk behaviors among injection drug users. Dans (Needle, R.H., Genser, S.G., and Trotter, R.T. II, eds). *Social Networks, Drug Abuse, and HIV Transmission*. NIDA Research Monograph 151. Rockville, MD: National Institute of Drug Abuse, 20-37.
- NEAIGUS, A., FRIEDMAN, S.R., JOSE, B., GOLDSTEIN, M.F., CURTIS, R., ILDEFONSO, G., et DES JARLAIS, D.C. (1996). High-risk personal networks and syringe sharing as risk factors for HIV infection among new drug injectors. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology* 11, 499-509.
- PALMER, E.M. (1985). *Graphical Evolution*. New York: Wiley.
- ROBINS, G.L. (1998). Personal attributes in inter-personal contexts: statistical models for individual characteristics and social relationships. Thèse de doctorat, University of Melbourne.
- ROSENBERGER, W.F. (1996). New directions in adaptive designs. *Statistical Science*, 11, 137-149.
- ROTHENBERG, R.B., WOODHOUSE, D.E., POTTERAT, J.J., MUTTH, S.Q., DARROW, W.W., et KLOVD AHL, A.S. (1995). Social networks in disease transmission: The Colorado Springs study. Dans (Needle, R.H., Genser, S.G., et Trotter, R.T. II, eds). *Social Networks, Drug Abuse, and HIV Transmission*. NIDA Research Monograph 151. Rockville, MD: National Institute of Drug Abuse, 3-19.
- RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- SCOTT, A.J. (1977). On the problem of randomization in survey sampling. *Sankhyā*, C, 39, 1-9.
- SCOTT, A.J., et SMITH, T.M.F. (1973). Survey design, symmetry, and posterior distributions. *Journal of the Royal Statistical Society, B*, 35, 57-60.
- SIRKEN, M.G. (1970). Household surveys with multiplicity. *Journal of the American Statistical Association*, 63, 257-266.
- SIRKEN, M.G. (1972a). Stratified sample surveys with multiplicity. *Journal of the American Statistical Association*, 67, 224-227.
- SIRKEN, M.G. (1972b). Variance components of multiplicity estimators. *Biometrics*, 28, 869-873.
- SIRKEN, M.G., et LEVY, P.S. (1974). Multiplicity estimation of proportions based on ratios of random variables. *Journal of the American Statistical Association*, 69, 68-73.
- SNUDDERS, T.A.B. (1992). Estimation on the basis of snowball samples: how to weight. *Bulletin de Méthodologie Sociologique*, 36, 59-70.
- SPREEN, M. (1992). Rare populations, hidden populations, and link-tracing designs: what and why? *Bulletin de Méthodologie Sociologique*, 36, 34-58.
- SPREEN, M. (1998). Sampling personal network structures: statistical inference in ego-graphs. Thèse de doctorat, University of Groningen.
- SPREEN, M., et ZWAAGSTRA, R. (1994). Personal network sampling, the network concept in studies of hidden populations. *International Sociology*, 9, 475-491.
- SUDMAN, S., SIRKEN, M.G., et COWAN, C.D. (1988). Sampling rare and elusive populations. *Science*, 240, 991-996.
- SUGDEN, R.A., et SMITH, T.M.F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, 71, 495-506.
- THOMPSON, S.K. (1990). Adaptive cluster sampling. *Journal of the American Statistical Association*, 85, 1050-1059.
- THOMPSON, S.K. (1997). Adaptive sampling in behavioral surveys. Dans (Harrison, L., et Hughes, A. eds.) *The Validity of Self-Reported Drug Use: Improving the Accuracy of Survey Estimates*. NIDA Research Monograph 167, Rockville, MD: National Institute of Drug Abuse, 296-319.
- THOMPSON, S.K., et SEBER, G.A.F. (1996). *Adaptive Sampling*. New York: Wiley.
- van MEETER, K.M. (1990). Methodological and design issues: techniques for assessing the representativeness of snowball samples. Dans (Lambert, E.Y. ed.) *The Collection and Interpretation of Data from Hidden Populations*. NIDA Monograph 98. Rockville, MD: National Institute on Drug Abuse, 31-43.
- WANG, Y.J., et WONG, G.Y. (1987). Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82, 8-19.
- WASSERMAN, S. (1980). Analyzing social networks as stochastic processes. *Journal of the American Statistical Association*, 75, 280-294.

Comme pour le calcul des estimations du maximum de vraisemblance elles-mêmes, le calcul de la matrice de l'information observée n'est pas influencé par le plan de désajustage de liens, puisque le plan est négligeable pour une vraisemblance fondée sur l'inférence. Cela s'oppose à l'information de Fisher prévue, dont la valeur est influencée par le plan de même que par le modèle de graphe, à moins qu'il ne s'agisse d'un plan conventionnel ne dépendant d'aucune valeur y ou x .

Pour un intervalle de prévision de niveau $(1 - \epsilon)$ pour une variable aléatoire telle que $n_1(s)$, une méthode consisterait à utiliser une région centrale de masse $(1 - \epsilon)$ de la fonction de vraisemblance de profil normalisée pour $n_1(s)$ (voir Björnsd 1990, 1996). Pour ce qui est du modèle symétrique, on peut facilement obtenir l'intervalle de prévision $(1 - \epsilon)$ pour $n_1(s)$, en calculant (19) pour $n_1(s) = 0, 1, 2, \dots$, jusqu'à ce que les valeurs calculées deviennent négligeables, puis en normalisant en divisant par le total cumulé $\sum_{n_1(s)=0}^L p$ et en utilisant les quantiles $\epsilon/2$ et $1 - \epsilon/2$ comme extrêmes de l'intervalle.

une variable aléatoire telle que $n_1^1(s)$, une méthode consisterait à utiliser une région centrale de masse $(1 - \varepsilon)$ de la fonction de vraisemblance de profil normalisée pour $n_1^1(s)$ (voir Bjørnstad 1990, 1996). Pour ce qui est du modèle symétrique, on peut facilement obtenir l'intervalle de prévision $(1 - \varepsilon)$ pour $n_1^1(s)$, en calculant (19) pour $n_1^1(s) = 0, 1, 2, \dots$, jusqu'à ce que les valeurs calculées deviennent négligéables, puis en normalisant en divisant par le total cumulé $\sum_{n_1^1(s)=0}^D$ et en utilisant les quantiles $\varepsilon/2$ et $1 - \varepsilon/2$ comme extrêmes de l'intervalle.

BIBLIOGRAPHIE

La présente recherche a pu bénéficier de l'appui de la National Science Foundation (DMS-9626102), des National Institutes of Health, National Institute on Drug Abuse (RO1 DA09872) et du Conseil suédois de la recherche en sciences humaines et sociales (HSFR F 0750/96).

BASU, D. (1969). Role of the sufficiency and likelihood principles in sample survey theory. *Samkhyā*, A 31, 441-454.

BIRNBAUM, Z.W., et SIRKEN, M.G. (1965). Design of sample surveys to estimate the prevalence of rare diseases: Three unbiased estimates. *Vital and Health Statistics*, 2, 11. Washington: Government Printing Office.

BJØRNSTAD, J.F. (1990). Predictive likelihood: A review. *Statistical Science*, 5, 242-265.

BJØRNSTAD, J.F. (1996). On the generalization of the likelihood function and the likelihood principle. *Journal of the American Statistical Association*, 91, 791-806.

DAWID, A.P., et DICKKEY, J.M. (1977). Likelihood and Bayesian inference from selectively reported data. *Journal of the American Statistical Association*, 72, 845-850.

EFRON, B., et HINKLEY, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information (avec discussion). *Biometrika*, 65, 457-487.

ERICKSON, B. (1979). Some problems of inference from chain data. *Sociological Methodology*, 10, 276-302

FIENBERG, S.E., MEYER, M.M., et WASSERMAN, S.S. (1985). Statistical analysis of multiple sociometric relations. *Journal of the American Statistical Association*, 80, 51-67.

FLOURNROY, N., et ROSENBERGER, W.F., Eds. (1995). *Adaptive Designs*. Hayward, CA: Institute of Mathematical Statistics.

FRANK, O. (1971). *Statistical Inference in Graphs*. Stockholm: Forssvarsforskningsanstalt.

FRANK, O. (1977a). Survey sampling in graphs. *Journal of Statistical Planning and Inference*, 1, 235-264.

FRANK, O. (1977b). Estimation of graph totals. *Scandinavian Journal of Statistics*, 4, 81-89.

FRANK, O. (1978a). Estimating the number of connected components in a graph by using a sampled subgraph. *Scandinavian Journal of Statistics*, 5, 177-188.

FRANK, O. (1978b). Sampling and estimation in large social networks. *Social Networks*, 1, 91-101.

FRANK, O. (1979a). Estimation of population totals by use of snowball samples. Dans *Perspectives on Social Network Research*, (eds. P.W. Holland et S. Leinhardt). New York: Academic Press, 319-347.

FRANK, O. (1979b). Moment properties of subgraph counts in stochastic graphs. *Annals of the New York Academy of Sciences*, 319, 207-218.

FRANK, O. (1981). A survey of statistical methods for graph analysis. *Sociological Methodology*, 110-155.

FRANK, O. (1988). Random sampling and social networks: a survey of various approaches. *Mathématiques, Informatique et Sciences humaines*, 26, 19-33.

FRANK, O. (1997). Composition and structure of social networks. *Mathématiques, Informatique et Sciences humaines*, 35, 11-23.

FRANK, O., et HARARY, F. (1982). Cluster inference by using transitivity indices in empirical graphs. *Journal of the American Statistical Association*, 77, 835-840.

FRANK, O., et SNIJDERS, T. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*, 10, 53-67.

FRANK, O., et STRAUSS, D. (1986). Markov graphs. *Journal of the American Statistical Association*, 81, 832-842.

FRIEDMAN, S.R., NEAIGUS, A., JOSE, B., CURTIS, R., GOLDSTEIN, M., ILDEFONSO, G., ROTHENBERG, R.B., et DESJARDIS, D.C. (1997). Sociometric risk networks and HIV risk. *American Journal of Public Health*. A paraitre.

GODAMBE, V.P. (1966). A new approach to sampling from finite populations. I. *Journal of the Royal Statistical Society B*, 28, 310-319.

GOODMAN, L.A. (1961). Snowball sampling. *Annals of Mathematical Statistics*, 32, 148-170.

GRANOVETTER, M. (1976). Network sampling: some first steps. *American Journal of Sociology*, 81, 1287-1303.

HOLLAND, P.W., LASKEY, K.B., et LEINHARDT, S. (1983). Stochastic block-models: First steps. *Social Networks*, 5, 109-137.

HOLLAND, P.W., et LEINHARDT, S. (1981). An exponential family of probability distributions for directed graphs (with discussion). *Journal of the American Statistical Association*, 76, 33-65.

FRANK, O. (1971). *Statistical Inference in Graphs*. Stockholm: Försvarshögskolans forskningsanstalt.

FRANK, O. (1977a). Survey sampling in graphs. *Journal of Statistical Planning and Inference*, 1, 235-264.

FRANK, O. (1977b). Estimation of graph totals. *Scandinavian Journal of Statistics*, 4, 81-89.

components in a graph by using a sampled Scandinavian Journal of Statistics, 5, 177-188.

FRANK, O. (1978b). Sampling and estimation in large social networks. *Social Networks*, 1, 91-101.

FRANK, O. (1979a). Estimation of population totals by use of snowball samples. Dans *Perspectives on Social Network Research*, (éds. P.W. Holland et S. Leinhardt). New York:

FRANK, O. (1979b). Moment properties of subgraph counts in stochastic graphs. *Annals of the New York Academy of Sciences*, 319, 207-218.

FRANK, O. (1981). A survey of statistical methods for graph analysis. *Sociological Methodology*, 110-155.

FRANK, O. (1988). Random sampling and social networks: a survey of various approaches. *Mathématiques, Informatique et Sciences humaines*, 26, 19-33.

FRANK, O. (1997). Composition and structure of social networks. *Mathématiques, Informatique et Sciences humaines*, 35, 11-23.

FRANK, O., et HARARY, F. (1982). Cluster inference by using transitivity indices in empirical graphs. *Journal of the American Statistical Association*, 77, 835-840.

FRANK, O., et al SNIJDEKRS, I. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*, 10, 53-67.

the American Statistical Association, 81, 832-842.

GOLDSTEIN, M., ILDEFONSO, G., ROTHENBERG, R.B., et
DESJARLAIS, D.C. (1997). Sociometric risk networks and HIV
risk. *American Journal of Public Health*. A paraitre.

GODAMBE, V.P. (1966). A new approach to sampling from finite populations. I. *Journal of the Royal Statistical Society B*, 28, 310-319.

GOODMAN, L.A. (1961). Snowball sampling. *Annals of Mathematical Statistics*, 32, 148-170.

American Journal of Sociology, 81, 1287-1303.

HOLLAND, P.W., LASKBY, K.B., et LEINHARDT, S. (1983). Stochastic block-models: First steps. *Social Networks*, 5, 109-137.

HOLLAND, P.W., et LEINHARDT, S. (1981). An exponential family of probability distributions for directed graphs (with discussion). *Journal of the American Statistical Association*, 76, 33-65.

À noter que les $n_j(s)$ pour $j = 0, 1$ sont compris en (15) uniquement dans les facteurs

$$\left(n(s) \prod_{j=0}^f V_j^{n_j(s)} \right)$$

où $V_j = \theta \prod_{i=0}^{j-1} \lambda_i^{n_i(s_0)}$. Puisque L est proportionnelle à une probabilité binomiale à paramètres $n(s)$ et $V_1 / (V_0 + V_1)$, il s'ensuit que le maximum de L pour $n_1(s)$ est obtenu pour $n_1(s)$ égal au nombre entier le plus proche de

$$\frac{n(s)V_1}{V_0 + V_1} + \frac{V_1}{2(V_0 + V_1)}$$

ou encore à l'un ou l'autre des nombres entiers les plus proches de ce nombre s'il y en a deux. En effet (voir par exemple Feller 1957, p. 140), le mode d'une distribution binomiale à paramètres (n, p) est le nombre entier dans l'intervalle $[(n+1)p - 1, (n+1)p]$ ou l'une ou l'autre des extrémités s'il s'agit de nombres entiers. Ainsi, le mode du nombre entier ou les nombres entiers les plus proches du point milieu de l'intervalle $(n+1)p - (1/2) = np + d - q/2$, où $q = 1 - p$.

Si les valeurs initiales des estimateurs de paramètre sont obtenues de la solution de (7) et substituées en V_j , une valeur prédite $n_1(s)$ est donnée comme ci-dessus. Si cette valeur prédite est insérée dans (16) et (17), on obtient de nouvelles estimations des paramètres, qui peuvent être substituées en V_j de façon à donner une nouvelle valeur prédite de $n_1(s)$, et ainsi de suite jusqu'à ce que les valeurs convergent vers la solution qui minimise (15). On peut également trouver la solution en calculant directement la vraisemblance (15) pour différentes valeurs de $n_1(s)$, en substituant les solutions obtenues de (16) et de (17) aux valeurs paramétriques.

5.2.1 Exemple: modèle symétrique

L'équation de vraisemblance prédictive (15) pour le modèle symétrique est

$$L[\theta, \beta; d, n_1(s)] = p(s | y^s, x^{s_0}) \left(\prod_{i=0}^f \theta_i^{n_i(s) + n_i'(s)} \right) \left(n(s) \prod_{j=1}^f V_j^{n_j(s)} \right) \times \left(\prod_{i=0}^f \beta_i^{n_i(s_0)} (1 - \beta_i)^{n_i'(s_0)} \right) \quad (18)$$

Soit $r^{kl} = r^{kl}(s_0, s)$, le nombre de paires de nœuds en $s_0 \times s$ comportant une valeur de nœud totale k et le nombre total de liens l . Pour le modèle symétrique, l ne peut avoir que la valeur 0, indiquant l'absence de liens entre les nœuds, ou 2, indiquant un lien symétrique. En particulier, $r^{02} = m^{0011}(s_0, s)$, $r^{12} = m^{0111}(s_0, s) + m^{1011}(s_0, s)$ et $r^{22} = m^{1111}(s_0, s)$ désignent le nombre de liens dans l'échantillon entre des nœuds de valeur totale k , pour $k = 0, 1, 2$, respectivement. Pour une telle notation, le dernier facteur en (18) peut s'écrire

5.3 Évaluation de l'exactitude des estimations

qui est une fonction de $n_1(s)$ seulement. Le prédicteur du maximum de vraisemblance de profil pour $n_1(s)$, facilement obtenu grâce à un calcul simple, est un nombre entier qui se situe entre 0 et $n(s)$ donnant la valeur la plus élevée de (19).

$$L^d[n_1(s); d] = p(s | y^s, x^{s_0}) \left(\prod_{i=0}^f \frac{N}{n_i(s) + n_i'(s)} \right) \left(n(s) \prod_{j=1}^f V_j^{n_j(s)} \right) \times \left(n(s) \prod_{k=0}^2 \left(\frac{c_k}{r^{k2}} \right)^{r^{k2}} \left(1 - \frac{c_k}{r^{k2}} \right)^{r_{k2}} \right)$$

et β_k aux paramètres en (18), ce qui donne profil pour $n_1(s)$ en substituant les valeurs maximisantes θ et β_k aux paramètres en (18), ce qui donne

On obtient la fonction de vraisemblance prédictive de $n_1(s)$. Pour une valeur donnée quelconque de $n_1(s)$, la vraisemblance est maximisée par $\theta_i = [n_i'(s) + n_j(s)] / \bar{N}$ pour $i = 0, 1, 2$. À noter que θ et les β_k sont des fonctions de la variable non observée $n_1(s)$.

Notons $c_k = c_k[n_1(s)]$ le nombre de paires de nœuds possibles en $s_0 \times U$ ayant une valeur totale k , de sorte que

$$c_k = r_{k*} + \sum_{i,j=0}^k n_i'(s_0) n_j(s)$$

$$\prod_{k=0}^2 \beta_k^{r_{k*}} (1 - \beta_k)^{r_{k*}} + \sum_{i,j=0}^k n_i'(s_0) n_j(s)$$

Pour des intervalles de confiance et d'autres formes d'inférence, on suggère l'inverse de l'information de Fisher observée $I(\phi)$, où ϕ est le vecteur des estimations du maximum de vraisemblance de paramètre et I est la matrice des secondes dérivées inversées de la fonction du log-likelihood du rapport de vraisemblance évaluée pour ces valeurs estimées. Le recours à l'information de Fisher observée plutôt que celle prévue pour évaluer l'exactitude d'une estimation est décrit dans Efron et Hinkley (1978). Plus récemment, Lindsay et Li (1997) ont soutenu que l'information observée fournit une meilleure évaluation de l'erreur de l'estimation qui est réalisée par opposition à celle qui est prévue. Pour préparer à l'aide de grands échantillons des approximations des propriétés des estimateurs de θ et λ , il importe d'adopter des hypothèses appropriées sur la façon dont λ dépend de N pour éviter la dégénérescence du modèle de graphe et de l'échantillon. Voir par exemple les résultats asymptotiques pour certains modèles de graphe simple dans Palmer (1985).

$$\theta_1 \beta_2 = r_{22} N_1 / (s_0)$$
$$({}^0s)_{1u}({}^2g - 1) = \frac{{}^0\theta/(s)_{0u} - N}{{}^1\theta/(s)_{1u} - N}$$
$$\begin{aligned} & \cdot \left(\begin{smallmatrix} \text{\tiny{0}} & \text{\tiny{0}} & \text{\tiny{0}} \\ \text{\tiny{f}} & \text{\tiny{u}} & \text{\tiny{0}} \end{smallmatrix} \text{\tiny{0}} \text{\tiny{f}} \text{\tiny{u}} \right) \left(\begin{smallmatrix} \text{\tiny{0}} & \text{\tiny{0}} & \text{\tiny{0}} \\ \text{\tiny{f}} & \text{\tiny{u}} & \text{\tiny{0}} \end{smallmatrix} \text{\tiny{0}} \text{\tiny{f}} \text{\tiny{u}} \right) \times \\ & \left(\begin{smallmatrix} \text{\tiny{0}} & \text{\tiny{0}} & \text{\tiny{0}} \\ \text{\tiny{f}} & \text{\tiny{u}} & \text{\tiny{0}} \end{smallmatrix} \text{\tiny{0}} \text{\tiny{f}} \text{\tiny{u}} \right) \left(\begin{smallmatrix} \text{\tiny{0}} & \text{\tiny{0}} & \text{\tiny{0}} \\ \text{\tiny{f}} & \text{\tiny{u}} & \text{\tiny{0}} \end{smallmatrix} \text{\tiny{0}} \text{\tiny{f}} \text{\tiny{u}} \right) \times \end{aligned}$$
$$42 / \binom{47}{2} = 0.039,$$

5.1.3 Modèle asymétrique

$$\frac{{}^0\theta/(s) {}^0u - N}{{}^1\theta/(s) {}^1u - N} = d$$
$$\frac{({}^0\theta/(s) {}^0u - N) \int_{\theta} d({}^0s) {}^1u + {}^0h u}{1 {}^1u} = \frac{h_p - 1}{h_p}$$
$$({}^{0s})^1u(x-1) = \frac{{}^0\theta/(s)^0u - N}{{}^1\theta/(s)^1u - N}$$

marqués, les équations à résoudre sont

$$\frac{\alpha}{\alpha} = \frac{1 - \alpha}{m^{011} + [n^1 \theta^1 N] + n^1 u^1 (s)^1} \cdot \frac{(s)^0 n^1 (s)^1}{m^{111}}$$

5.2 Vraisemblance prédictive pour le total des valeurs de nœud non observées

$$\begin{pmatrix} \cdot & \cdot & 0 & \dot{f}_i \\ (\underline{s})^t u & (0s)^t u & \gamma \dot{\Pi} & \cdot \end{pmatrix} \begin{pmatrix} \cdot & \cdot & \gamma \dot{f}_i \\ (1s \ 0s)^t \gamma \dot{f}_i & \gamma \dot{\Pi} & \dot{D} \end{pmatrix} \times$$

(15) en remplaçant les estimations θ et λ qui maximisent la vraisemblance (marginale) (5). La valeur de $n_1(\hat{s})$ maximisant la vraisemblance estimative serait le prédicteur du maximum de vraisemblance estimatif de $n_1(\hat{s})$. Les méthodes de vraisemblance estimative tendent à produire des prévisions raisonnablement d'un point unique dans de nombreux cas, mais elles sont moins utiles pour la prévision

en (15) et de maximiser cette vraisemblance estimative relativement à $n^1(s)$, la vraisemblance (15) est maintenant maximisée simultanément pour ce qui est de chaque paramètre et de $n^1(s)$. Cela signifie que, pour chaque valeur de $n^1(s)$, il existe des valeurs de paramètre $\theta[n^1(s)]$ et $\gamma^{(k)}[n^1(s)]$ qui maximisent (15) relativement à γ . Le fait de substituer ces valeurs en (15) permet de définir la vraisemblance de profil $\bar{L}(s)$ pour $n^1(s)$.

$$(91) \quad \frac{N}{(s)'u + (s)'u} = \theta$$

Pour les paramètres qui restent, on utilise $d \log L / da^j$ et $d \log L / d\beta^k$ tirées de (8) et (9), les dérivées partielles étant maintenant données par

$$(L1) \cdot \frac{\gamma_{00}}{(\underline{s})^f (u)^{0_s} u} (\kappa - 1) + \frac{\gamma_{01}}{(\underline{s})^f (u)^{0_s} u} + \frac{\gamma_{11}}{(\underline{s})^{0_s} (u)^{0_s} u} = \frac{\gamma_e}{T \log e}$$

également que $\lambda_{ji0}^* = 1 - \alpha_{ji}^*$ et que $\lambda_{ji1}^* = \alpha_{ji}^*$ peut être remplacé pour simplifier la vraisemblance.

5.1.1 Équations de vraisemblance estimative

Les estimateurs du maximum de vraisemblance pour les paramètres θ_1, α_{ij} et β_k sont obtenus comme solution commune des équations

$$\frac{\partial \theta_1}{\partial \log L} = \frac{\partial \alpha_{ij}}{\partial \log L} = \frac{\partial \beta_k}{\partial \log L} = 0 \quad (7)$$

pour $i = 0, 1, j = 0, 1, k = 0, 2$. Si l'on différencie le logarithme de la vraisemblance (5) relativement à θ_1 et si l'on règle à zéro, on a

$$\frac{\partial \theta_1}{\partial \log L} = \frac{\partial \theta_1}{\partial \log L} - \frac{\partial \theta_1}{\partial \log L} = 0$$

où les dérivées partielles sont données par

$$\frac{\partial \theta_k}{\partial \log L} = \frac{\theta_k}{n(s)} + n(s) \frac{\sum_j \theta_j \prod_i \lambda_{ij0}^*}{\prod_i \lambda_{ij0}^*}$$

pour $k = 0, 1$.

De plus,

$$\frac{\partial \alpha_{ij}}{\partial \log L} = \frac{\partial \alpha_{ij}}{\partial \log L} + \frac{\partial \alpha_{ij}}{\partial \log L} - \frac{\partial \alpha_{ij}}{\partial \log L} = 0 \quad (8)$$

et

$$\frac{\partial \beta_k}{\partial \log L} = \sum_{i,j=0,1} \left(\frac{\partial \beta_k}{\partial \log L} + \frac{\partial \beta_k}{\partial \log L} - \frac{\partial \beta_k}{\partial \log L} \right) \quad (9)$$

où les dérivées partielles sont données par

$$\frac{\partial \log L}{\partial \lambda_{ijk}^*} = \frac{\lambda_{ijk}^*}{m_{ijk}^*(s_0, s_1)} + \frac{\lambda_{ijk}^*}{m_{ijk}^*(s_0, s_1)}$$

$$+ (1 - k)n(s) \frac{\sum_j \theta_j \prod_i \lambda_{ij0}^*}{\prod_i \lambda_{ij0}^*}$$

Il est commode d'écrire l'équation de vraisemblance pour θ_1 sous la forme

$$\frac{\theta_1}{n_1(s)} - \frac{\theta_0}{n_0(s)} + \frac{\theta_1 p + \theta_0}{n(s)(p - 1)} = 0 \quad (10)$$

où

$$p = \prod_{i=0,1} \left(\frac{\lambda_{i10}^*}{\lambda_{i00}^*} \right) = \prod_{i=0,1} \left(\frac{1 - \alpha_{i1}^*}{1 - \alpha_{i0}^*} \right)$$

À noter que $p = p_0 = p_1$, où $p_i = (1 - \alpha_{i1}^*) / (1 - \alpha_{i0}^*)$ est le rapport entre les probabilités d'une absence d'arc allant d'un nœud i à un nœud positif et à un nœud zéro, respectivement.

Admettant $m_{ijk}^*(s_0, s_1) = r_{i+j,k+1}$ on obtient des estimateurs du maximum de vraisemblance comme solution

$$\frac{n_1(s)}{n_0(s)} - \frac{\theta_1}{n_0(s)} - \frac{\theta_0}{n(s)(1 - p)} = 0 \quad (11)$$

$$\frac{\beta_0}{r_{02}} - \frac{1 - \beta_0}{r_{00}} - \frac{(1 - \beta_0)(\theta_0 + p\theta_1)}{n(s)n_0(s)\theta_0} = 0 \quad (12)$$

$$\frac{\beta_1}{r_{12}} - \frac{1 - \beta_1}{r_{10}} - \frac{(1 - \beta_1)(\theta_0 + p\theta_1)}{n(s)[n_1(s)\theta_0 + n_0(s)p\theta_1]} = 0 \quad (13)$$

$$\frac{\beta_2}{r_{22}} - \frac{1 - \beta_2}{r_{20}} - \frac{(1 - \beta_2)(\theta_0 + p\theta_1)}{n(s)n_1(s)p\theta_1} = 0 \quad (14)$$

Les modèles symétriques comportent $\lambda_{ijk}^* = 0$ pour $k \neq l$ de sorte que les arcs sont toujours mutuels ou que, ce qui revient au même, ils peuvent être considérés comme des contours non dirigés. Le modèle symétrique complet comporte des paramètres $\lambda_{ijk}^* = \lambda_{jik}^*$ pour $i, j, k = 0, 1$, avec $\lambda_{j00}^* + \lambda_{j11}^* = 1$. Ici $\lambda_{ji1}^* = \beta_{i+j}^* = \alpha_{ij}^*$ et

$$p = \prod_{i=0,1} \left(\frac{1 - \beta_{i+1}^*}{1 - \beta_i^*} \right)$$

5.1.2 Modèle symétrique

On peut établir une interprétation de l'influence de la structure du graphe sur l'estimation de θ_1 en considérant les paramètres de graphe α – et donc p – comme fixes. La proportion de nœuds positifs dans l'échantillon est notée $\theta_c^* = n_1(s)/n(s)$. C'est la l'estimateur conventionnel ou naïf de θ_1 , d'après la proportion de nœuds positifs dans l'échantillon. Si $p = 1$, l'estimateur du maximum de vraisemblance $\hat{\theta}_1$ serait $\hat{\theta}_c^*$. Si $p < 1$, l'estimateur du maximum de vraisemblance $\hat{\theta}_1$ serait inférieur à $\hat{\theta}_c^*$, et si $p > 1$, $\hat{\theta}_1 > \hat{\theta}_c^*$. En particulier, $\alpha_{i1}^* = \alpha_{i0}^*$ pour $i = 0, 1$ suppose $p = 1$ et l'estimateur du maximum de vraisemblance est $\hat{\theta}_1 = \hat{\theta}_c^*$.

Considérons par exemple le cas dans lequel, pour une valeur quelconque de y^* , un lien entre le nœud u et le nœud v est plus vraisemblable lorsque $y^* = 1$ que lorsque $y^* = 0$, de sorte que $\alpha_{i1}^* > \alpha_{i0}^*$ pour $i = 0, 1$. Dès lors $(1 - \alpha_{i1}^*) / (1 - \alpha_{i0}^*) < 1$ pour $i = 0, 1$, de sorte que $p < 1$ et l'estimateur conventionnel $\hat{\theta}_c^*$. On pourrait affirmer que le plan à dépistage de liens entraîne les chercheurs vers une proportion élevée de nœuds positifs qui n'est pas représentative et que l'estimateur du maximum de vraisemblance établit l'ajustement.

Dans des cas particuliers, on pourra fixer certains λ_{ijk}^* à zéro et il faudra modifier les équations de vraisemblance en conséquence. Quelques cas particuliers sont examinés ci-dessous.

Pour ce qui est des y_i ,

$$\prod_{i=1}^I \prod_{j=0}^{I-1} \prod_{k=0}^{I-1} \prod_{l=0}^{I-1} \lambda_{ijkl}^{y_{ijkl}} = \left(\prod_{j=0}^{I-1} \prod_{k=0}^{I-1} \prod_{l=0}^{I-1} R_{jkl}^{y_{jkl}} \right) (\gamma^{R_{11}}).$$

où les R sont des dénombrements de dyades correspondant au schéma du tableau I. Autrement dit, $R_{00} = M_{0000}$,

$R_{01} = M_{0001} + M_{0010}$, $R_{02} = M_{0011} + M_{0100}$, $R_{10} = M_{0100} + M_{1001}$, $R_{11} = M_{0101} + M_{1010}$, $R_{12} = M_{0110} + M_{1011}$, $R_{20} = M_{1100} + M_{1101}$, $R_{21} = M_{1101} + M_{1110}$, $R_{22} = M_{1110} + M_{1111}$. À noter que

$R_{11}^{(R_{11})}$ représente le nombre de dyades comportant un arc allant d'un nœud non marqué (marqué) à un nœud marqué (non marqué) seulement. À noter également que, sauf pour $(ij) = (11)$, R_{ij} est le nombre de dyades pour i nœuds marqués comportant j arcs.

Les estimateurs du maximum de vraisemblance pour le graphe entier comme données sont les proportions $\hat{\theta}_i = N_i/N$, $\hat{\gamma}_{ij} = R_{ij}/R_1$, et $\hat{\gamma}_{11} = R_{11}/R_1$, où $R_0 = N_0(N_0 - 1)/2$, $R_1 = N_0 N_1$ et $R_2 = N_1(N_1 - 1)/2$. Pour ce qui est des λ_i , cela signifie que $\lambda_{ijkl} = R_{11}^{(R_{11})}/R_1$ si $(ijkl) = (0110)$ ou (1001) et que $\lambda_{ijkl} = R_{1-j, k-l}/R_{1-j}$ autrement.

5. INFÉRENCE TIRÉE DE PLANS À DÉPISTAGE DE LIENS

5.1 Estimation de paramètres de modèle de graphe

Considérons l'un ou l'autre des plans à dépistage de liens pour lesquels on tire un échantillon initial ou à plusieurs vagues, quitte à suivre les liens établis à partir des nœuds en s_0 pour ajouter l'ensemble s_1 des nœuds ne figurant pas en s_0 qui sont contigus après les nœuds en s_0 . Les données sont $d = (s, Y^s, X^{s_0 U})$, de sorte que le plan dépend des valeurs y et x uniquement par l'entremise de celles des données et est donc négligeable.

Suivant le modèle de graphe décrit à la section précédente, la vraisemblance pour les données d'échantillon données par l'équation (2) de la section 2 est dans ce cas-ci

$$L(\theta, \lambda, d) = P(s | y^s, x^{s_0 U}) = \sum_{n=1}^N \left(\prod_{i=1}^n \theta_{y_i} \right) \left(\prod_{i=1}^{n > n'} \lambda_{y_i x^{s_0 U} y_i x^{s_0 U}} \right)$$

où la somme englobe toutes les valeurs y^n et $x^{s_0 U}$ qui ne sont pas fixées par les données de l'échantillon.

Comme pour la notation des chiffres de population de la section précédente, $n_1(a)$ désigne le nombre de nœuds

$n \in a$ comportant $y^n = i$ pour des sous-ensembles arbitraires $a \subset U$. Soit $m_{ijl}^{(a, b)}$, le nombre de paires de nœuds (u, v) tel que $u \in a$, $v \in b$, $(y^u, y^v, x^{uv}, x^{vu}) = (ijkl)$ et $n < v$ si u aussi bien que v appartenant à $a \cap b$. Un indice remplacé par un point représente une sommation pour cet indice. Ainsi, en conformité avec les plans à dépistage de liens décrits à la section 3, on observe uniquement

$$m_{ijkl}^{(s_0, s_1)}, \text{ et non pas } m_{ijkl}^{(s_0, s_1)}.$$

Pour les données de l'un ou l'autre plan à dépistage de liens décrit à la section 3, la fonction de vraisemblance est

$$L(\theta, \lambda, d) = P(s | y^s, x^{s_0 U}) = \left(\prod_{i=1}^n \theta_{y_i} \right) \left(\prod_{i=1}^{n'} \lambda_{y_i x^{s_0 U} y_i x^{s_0 U}} \right) \times \left(\prod_{i=1}^{n'} \prod_{j=0}^{I-1} \prod_{k=0}^{I-1} \prod_{l=0}^{I-1} \lambda_{ijkl}^{y_{ijkl}} \right).$$

Pour les plans à dépistage de liens dans lesquels on suit tous les liens au lieu d'un sous-échantillon à partir de l'échantillon initial, on a zéro pour tous les éléments de la sous-matrice X^{s_0} et $m_{i(s_0), v}^{(s_0, s_1)} = n_i(s_0)$ pour $v \in s_1$, ce qui réduit la fonction de vraisemblance à

$$L(\theta, \lambda, d) = P(s | y^s, x^{s_0 U}) = \left(\prod_{i=1}^n \theta_{y_i} \right) \left(\prod_{i=1}^{n'} \lambda_{y_i x^{s_0 U} y_i x^{s_0 U}} \right) \times \left(\prod_{i=1}^{n'} \prod_{j=0}^{I-1} \prod_{k=0}^{I-1} \prod_{l=0}^{I-1} \lambda_{ijkl}^{y_{ijkl}} \right).$$

Le facteur $\prod_{i=1}^n \theta_{y_i}$ donne la probabilité des valeurs de nœud observées dans l'échantillon. Le facteur $\prod_{i=1}^{n'} \lambda_{ijkl}^{y_{ijkl}}$ donne la probabilité des types de dyades observés en $s_0 \times s_1$. Puisque $x^{s_0 U}$ mais non pas x^{s_0} est observé, pour $n \in s_0$ et $v \in s_1$, la probabilité marginale que

$x^{uv} = k$ étant donné $y^n = i$ et $y^v = j$ est $\lambda_{ijkl}^{y_{ijkl}}$. Le dernier facteur de (5), entre crochets, donne la probabilité qu'il n'y ait pas d'arcs allant de l'échantillon initial à s_1 . Pour un nœud v parmi les $n(s_1)$ nœuds hors de l'échantillon, θ_j est la probabilité que $y^v = j$. Pour l'un ou l'autre des nœuds $n \in s_0$ de l'échantillon $n(s_0)$ comportant $y^n = i$, la probabilité conditionnelle qu'il n'y ait aucun lien avec v , c'est-à-dire que $x^{nv} = 0$, $\lambda_{ij00}^{y_{ij00}}$.

De façon plus formelle, on peut obtenir le terme entre crochets par conditionnement pour le nombre $n(s_1)$ de nœuds de type j en s_1 . Suivant $n_j(s_1)$, la probabilité d'une valeur zéro pour tous les indicateurs de lien allant de s_0 à s_1 est obtenue comme suit. Entre les nœuds $n(s_0)$ de type i en s_0 et les nœuds $n(s_1)$ de type j en s_1 , la probabilité d'une valeur zéro pour tous les liens est $\lambda_{ij00}^{y_{ij00}}$. Si l'on utilise la loi binomiale de $n_j(s_1)$ et la loi des probabilités totales, la probabilité d'une valeur zéro pour tous les liens allant de s_0 à s_1 , étant donné y^s , est

$$\sum_{n=0}^{n_1(s_1)} \binom{n_1(s_1)}{n} \left(\prod_{i=1}^n \theta_{y_i} \right) \left(\prod_{i=1}^{n'} \prod_{j=0}^{I-1} \prod_{k=0}^{I-1} \prod_{l=0}^{I-1} \lambda_{ijkl}^{y_{ijkl}} \right) = \sum_{i=1}^I \theta_i \prod_{j=0}^{I-1} \prod_{k=0}^{I-1} \prod_{l=0}^{I-1} \lambda_{ij00}^{y_{ij00}}.$$

Pour le plan à vague terminée, les expressions de vraisemblance ci-dessus sont simplifiées puisque les termes comportant ces termes de valeur un. À noter

modèle, il importe d'avoir des outils diagnostiques en vue des évaluations et des comparaisons avec d'autres modèles. Ainsi, pour le modèle à deux blocs utilisé ici, l'indépendance conditionnelle des dyades peut être vérifiée en comptant les paires de dyades de type différent au sein des blocs et entre ceux-ci. Dans chaque bloc on trouve trois types de dyades et six types de paires de dyades. Entre les deux blocs il y a quatre types de dyades et dix types de paires de dyades. Une fonction de la qualité de l'ajustement de Pearson entre des comptes observés et prévus des 22 types de paires de dyades au sein des blocs et entre ceux-ci comporte une distribution chi carré asymptotique à 12 degrés de liberté pour l'hypothèse d'une indépendance conditionnelle des dyades. On trouvera une discussion des essais de qualité de l'ajustement pour des modèles de graphe dans Holland et Leinhardt (1981) et dans Frank et Strauss (1986), et il y a lieu de poursuivre la recherche en ce sens, surtout pour ce qui est des données d'échantillon tirées de plans à dépiçages de liens.

Dans le modèle hypothétique, les variables de nœud X^1, \dots, X^N sont des variables aléatoires de Bernoulli indépendantes et distribuées de façon identique à probabilités $P(X^i = ?) = \theta_i$, pour $i = 0, 1$, avec $\theta_0 + \theta_1 = 1$. Suivant les valeurs de nœud X^1, \dots, X^N , les dyades (X^{nv}, X^{vn}) sont indépendantes, pour $1 \leq n < v \leq N$, et comportent une distribution conditionnelle donnée par $P[(X^{nv}, X^{vn}) =$

Tableau I

$(x^n, x^{n-1}x^n)$	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
$(0,0)$				
$(0,1)$				
$(1,0)$				
$(1,1)$				

Compte tenu de ces restrictions, il est commode d'introduire la notation

Diagram illustrating the four basis states for two qubits, labeled $|1'1\rangle$, $|0'1\rangle$, $|1'0\rangle$, and $|0'0\rangle$. Each state is represented by a 2x2 grid of dots. The connections between dots represent the state:

- $|1'1\rangle$: Horizontal connections between the top and bottom dots in each column.
- $|0'1\rangle$: Diagonal connections between the top-left and bottom-right dots, and the top-right and bottom-left dots.
- $|1'0\rangle$: Vertical connections between the top and bottom dots in each column.
- $|0'0\rangle$: Horizontal connections between the top and bottom dots in each column.

Below the diagrams is the expression $(n A_X e^{i n X})$.

$$\left. \begin{array}{l} \gamma'_{ij, k+l} \text{ si } (ijkl) = (0110) \text{ ou } (1001), \\ \gamma_{ij, k+l} \text{ autrement} \end{array} \right\} = \chi_{ijkl}$$

on $\gamma_{00} + 2\gamma_{01} + \gamma_{02} = 1$, $\gamma_{10} + \gamma_{11} + \gamma_{12} = 1$ et $\gamma_{20} + 2\gamma_{21} + \gamma_{22} = 1$. Nous pouvons interpréter γ_{ij}^1 et γ_{ij}^2 comme la probabilité de dyades ayant un arc allant d'un nœud non marqué à un nœud marqué seulement et allant d'un nœud marqué à un nœud non marqué seulement, respectivement. De plus, pour $(ij) \neq (11)$, γ_{ij}^2 est la probabilité d'une dyade ayant i arcs pour i nœuds marqués et $2-i$ nœuds non marqués.

Il est également commode de noter $\lambda_{ji} = \sum_{k=1}^n \lambda_{ijk}$, $\alpha_j = \alpha_{jj}$ et $\beta_{ji} = \beta_{ji}$ pour $i = 0, 1$ et $j = 0, 1$. Ici α_j est la probabilité d'un arc allant d'un nœud de valeur j à un nœud de valeur j , et β_j est la probabilité d'arcs mutuels entre k nœuds marqués.

Soit N_j , le nombre total de nœuds de valeur j dans le graphe, pour $j = 0, 1$, de sorte que $N_0 + N_1 = N$. Soit M_{ijk} , le nombre total de dyades de type (ijk) , c'est-à-dire le nombre total de paires de nœuds ordonnées (u, v) , avec $u < v$, de sorte que $(X^u, X^v) = (X^u, X^v)$.

$$(3) \quad \left(\begin{array}{ccccc} \rho_{ij}^{f_i} & 0=i & 0=j & 0=f & 0=i \\ \rho_{ij}^{f_i} & \prod_1 & \prod_1 & \prod_1 & \prod_1 \end{array} \right) \left(\begin{array}{c} 0=i \\ \theta \prod_1 \\ i_N \prod_1 \end{array} \right) = (\mathbf{x}, \mathbf{y}, \mathbf{z}, \theta, \tau) L$$

3.4 Dépistage de liens adaptable aux valeurs de nœud

Considérons un plan dans lequel la décision de suivre les liens depuis le nœud n dépend de la valeur de nœud y^n .

Dans une étude de l'usage de drogues injectables, par exemple, l'échantillon initial peut comporter à la fois des utilisateurs ($Y^n = 1$) et des non-utilisateurs ($Y^n = 0$). Si les chercheurs décident de suivre les liens sociaux uniquement à partir des utilisateurs, le plan dépend, du point de vue de l'adaptation, des valeurs y de nœud de même que des liens. De même, dans une étude des maladies transmises sexuellement, les chercheurs peuvent avoir à suivre les liens sexuels ou sociaux plus fréquemment à partir de répondants infectés qu'à partir de ceux qui ne sont pas infectés. Le plan peut alors s'écrire sous la forme $p(s|y^s, x_{s0}^U)$, puisque la procédure de sélection dépend à la fois des valeurs de nœud et des valeurs de lien. Si les données comportent toutes les valeurs dont dépend le plan, c'est-à-dire $d = (s, y^s, x_{s0}^U)$, le plan est négligeable et l'inférence du maximum de vraisemblance est simplifiée comme le décrivent les sections ci-dessous.

3.5 Dépistage d'un sous-échantillon seulement de liens de l'échantillon

Les plans décrits ci-dessus se laissent généraliser en procédures dans lesquelles on suit un échantillon seulement des liens établis à partir du nœud n en s_0 . Les exemples englobent le plan en «marche aléatoire» de Klov Dahl (1989) et la généralisation des plans en boule de neige décrite dans Snijders (1992). Pour ce qui est du plan de marche aléa-

toire, un répondant initial est prié de fournir le nom de plusieurs contacts sociaux. Un de ces contacts est choisi au hasard, interviewé et prié de fournir à son tour le nom de plusieurs contacts, un de ceux-ci étant choisi au hasard, et ainsi de suite. Concrètement, un cul-de-sac peut se produire lorsque le répondant ne fournit aucun nom, ou que les contacts mentionnés se trouvent déjà dans l'échantillon. Dans un tel cas, le chercheur doit revenir en arrière et suivre des pistes de répondants antérieurs, ou trouver un nouveau répondant initial.

Pour de tels plans à dépistage de liens par sous-échantillonnage, la procédure de tirage de l'échantillon, bien que complexe du point de vue de la probabilité du plan, dépend uniquement de valeurs de l'échantillon et de liens établis à partir de l'échantillon. Encore une fois, nous supposons que l'échantillon initial est obtenu à l'aide d'une procédure négligeable quelconque. Soit $s_0 = s_{00} \cup s_{01} \cup s_{02} \cup \dots \cup s_{0k}$, constitue de toutes les vagues dont quelques liens au moins sont suivis. Ainsi, s_{01} comporte les nœuds non inclus antérieurement et obtenus en suivant un sous-échantillon des liens établis à partir de nœuds de l'échantillon initial s_{00} , s_{02} comportant les nœuds non inclus antérieurement et obtenus en suivant un sous-échantillon des liens établis à partir de la vague et ainsi de suite. Aucun lien n est suivi à partir de la vague

4. MODÈLE DE GRAPHE COMPORTANT DES LIENS ASSOCIÉS À DES VALEURS DE NŒUD

La méthode fondée sur la vraisemblance décrite à la

section 2 en fonction de données d'échantillon tirées de plans à dépistage de liens décrits à la section 3 sera maintenant illustrée à l'aide d'une catégorie de modèles s'appuyant sur une indépendance conditionnelle entre les données comme dans les modèles de contact de Frank (1979a) et de Wellman et coll. (1991). Suivant les valeurs de nœud, on suppose l'indépendance entre dyades, la distribution des liens entre paires de nœuds dépendant de la valeur du nœud. Ainsi, inconditionnellement ces modèles comportent une dépendance entre dyades à cause de la dépendance à l'égard des valeurs de nœud. Dans les modèles de Holland et Leinhardt (1981), on suppose que les dyades sont indépendantes, mais qu'elles comportent des distributions qui dépendent de paramètres de nœud fixes. Wasserman (1980) a également supposé l'indépendance des dyades dans la modélisation du changement d'un graphe dans le temps comme processus stochastiques de

Snijders (1992). Pour ce qui est du plan de marche aléa-
toire, un répondant initial est prié de fournir le nom de
plusieurs contacts sociaux. Un de ces contacts est choisi au
hasard, interviewé et prié de fournir à son tour le nom de
plusieurs contacts, un de ceux-ci étant choisi au hasard, et
ainsi de suite. Concrètement, un cul-de-sac peut se produire
lorsque le répondant ne fournit aucun nom, ou que les
contacts mentionnés se trouvent déjà dans l'échantillon.
Dans un tel cas, le chercheur doit revenir en arrière et suivre
des pistes de répondants antérieurs, ou trouver un nouveau
répondant initial.

Considérons un plan dans lequel la décision de suivre les liens depuis le nœud n dépend de la valeur de nœud y^n . Dans une étude de l'usage de drogues injectables, par exemple, l'échantillon initial peut comporter à la fois des utilisateurs ($Y^n = 1$) et des non-utilisateurs ($Y^n = 0$). Si les chercheurs décident de suivre les liens sociaux uniquement à partir des utilisateurs, le plan dépend, du point de vue de l'adaptation, des valeurs y de nœud de même que des liens. De même, dans une étude des maladies transmises sexuellement, les chercheurs peuvent avoir à suivre les liens sexuels ou sociaux plus fréquemment à partir de répondants infectés qu'à partir de ceux qui ne sont pas infectés. Le plan peut alors s'écrire sous la forme $p(s|y^s, x_{s0}^U)$, puisque la procédure de sélection dépend à la fois des valeurs de nœud et des valeurs de lien. Si les données comportent toutes les valeurs dont dépend le plan, c'est-à-dire $d = (s, y^s, x_{s0}^U)$, le plan est négligeable et l'inférence du maximum de vraisemblance est simplifiée comme le décrivent les sections ci-dessous.

3.6 Données provenant de plans à dépistage de liens

Pour l'un ou l'autre des plans à dépistage de liens comportant une ou plusieurs vagues décrites ci-dessus, il est bien important en pratique de savoir quelles données sont enregistrées. Si les données englobent uniquement les étiqettes des nœuds de l'échantillon, les valeurs y des nœuds de l'échantillon et les indicateurs d'arc des paires de nœuds de l'échantillon, c'est-à-dire que $d = (s, y^s, x_{s0}^U)$, le plan n'est pas négligeable et doit être intégré à la vraisemblance, ce qui risque de compliquer l'analyse.

Considérons également une étude dans laquelle des liens sociaux sont utilisés dans le plan afin d'établir l'échantillon, mais dans laquelle seules les caractéristiques de nœud (valeurs y) et non pas les relations sont enregistrées, de sorte que les données sont $d = (s, y^s)$. Le plan alors n'est pas négligeable.

Si, par contre, les données tirées du plan à dépistage de liens englobent non seulement les liaisons au sein de l'échantillon, mais également les liaisons qui en sortent (ou l'absence de telles liaisons), pour toutes les vagues sauf la dernière, avec le reste du graphe, c'est-à-dire que $d = (s, y^s, x_{s0}^U)$, le plan dépend uniquement de valeurs de graphe parmi les données et on n'en tient pas compte pour la vraisemblance.

La méthode fondée sur la vraisemblance décrite à la section 2 en fonction de données d'échantillon tirées de plans à dépistage de liens décrits à la section 3 sera maintenant illustrée à l'aide d'une catégorie de modèles s'appuyant sur une indépendance conditionnelle entre les données comme dans les modèles de contact de Frank (1979a) et de Wellman et coll. (1991). Suivant les valeurs de nœud, on suppose l'indépendance entre dyades, la distribution des liens entre paires de nœuds dépendant de la valeur du nœud. Ainsi, inconditionnellement ces modèles comportent une dépendance entre dyades à cause de la dépendance à l'égard des valeurs de nœud. Dans les modèles de Holland et Leinhardt (1981), on suppose que les dyades sont indépendantes, mais qu'elles comportent des distributions qui dépendent de paramètres de nœud fixes. Wasserman (1980) a également supposé l'indépendance des dyades dans la modélisation du changement d'un graphe dans le temps comme processus stochastiques de

3.2 Échantillons à plusieurs vagues

Considérons un échantillon en boule de neige comportant $k + 1$ vagues après l'échantillon initial. L'échantillon est noté $s = s_0 \cup s_1$ avec $s_0 = s_0^0 \cup s_0^1 \cup \dots \cup s_0^k$. Un échantillon initial s_0^0 est tiré à l'aide de tout plan qui est négligable pour ce qui de la vraisemblance. On suit les liens, et chaque nœud comportant un arc pour tout nœud s_0^0 qui n'est pas encore dans l'échantillon est ajouté de façon à former l'échantillon de la première vague s_0^1 . Autrement dit, $s_0^1 = \{v : x^{uv} = 1 \text{ pour certains } u \in s_0^0, v \notin s_0^0\}$. On suit alors les liens en s_0^1 pour obtenir l'échantillon de la deuxième vague $s_0^2 = \{v : x^{uv} = 1 \text{ pour certains } u \in s_0^1, v \notin s_0^0 \cup s_0^1\}$. Enfin, l'échantillon de la vague $(k + 1)$, noté simplement s_1 , est ajouté en suivant les liens de l'échantillon s_0^k de la k -ième vague. Autrement dit, $s_1 = \{v : x^{uv} = 1 \text{ pour certains } u \in s_0^k, v \notin s_0\}$. On ne suit aucun lien de s_1 .

Si $s_0^j = \emptyset$ pour tout $j < k$ on cesse d'échantillonner, de sorte que le nombre de vagues ajoutées est inférieur à k si, en un point quelconque, aucun lien n'est établi entre l'échantillon courant et des nœuds non échantillonnés.

Les données englobent des ensembles d'étiquettes de nœuds différentes vagues de l'échantillon et les paires de nœuds ordonnées entre s_0 et U , la séquence des valeurs de nœud y_s pour tous les nœuds de l'échantillon, de même que les variables indicatrices de lien $x_{s_0 U}$ entre s_0 et l'ensemble U de nœuds du graphe. Les données englobent donc les données de sous-graphe pour s_0 , c'est-à-dire $(s_0, y_{s_0}, x_{s_0 U})$, de même que les valeurs de nœud y_{s_1} pour les nœuds de la vague définitive s_1 , les indicateurs de lien $x_{s_0 s_1}$ entre s_0 et s_1 , de même que les indicateurs de lien $x_{s_0 s_1}$ entre les nœuds en s_0 et les nœuds qui ne sont pas dans l'échantillon.

Puisque le plan ne dépend d'aucune valeur y ou x à l'extérieur des données, ni d'aucun paramètre du modèle de graphe, il est négligable et la structure des données est exactement la même pour le plan en boule de neige de la vague $(k + 1)$ et pour le plan en boule de neige de la vague 1 ; avec la notation que nous avons utilisée, les formules de vraisemblance et d'estimation restent inchangées pour le plan plus général.

3.3 Plans à vague terminée

Pour un échantillon en boule de neige terminé, l'ajout de vagues se poursuit jusqu'à ce que l'échantillon ne comporte plus de liens. Le nombre de vagues terminées K est alors une variable aléatoire et $s_0, K + 1 = s_1$ est le premier ensemble vide de la séquence (s_0^0, s_0^1, \dots) . Les données sont $d = (s_0, y_{s_0}, x_{s_0 U})$ ou l'équivalent $(s_0, y_{s_0}, x_{s_0 s_0^1}, x_{s_0 s_0^2}, \dots)$. L'intérêt peut alors se poursuivre avec les mêmes formules de vraisemblance et d'estimation, mais aussi avec une simplification: les données ne contiennent aucun ensemble s_1 pour lequel y_{s_1} et $x_{s_0 s_1}$ sont connus mais pour lequel on ne connaît pas de liens.

$s^{(2)}$ comporte une relation fonctionnelle déterministe avec l'échantillon de nœuds $s^{(1)}$, la notation en exposant est laissée de côté, et l'échantillon de nœuds définitif $s^{(1)}$ est noté simplement s . Les méthodes de vraisemblance simple décrites dans le présent exposé s'appliquent à tout un choix de plans à dépistage de liens négligables, y compris ceux qui sont décrits dans la présente section. Il y a lieu de poursuivre la recherche sur les méthodes pour des plans non négligables, y compris ceux qui comportent une sélection non négligable de l'échantillon initial. Les méthodes de traitement des erreurs non dues à l'échantillonnage, par exemple les erreurs de non-réponse et de déclaration, à l'aide de plans à dépistage de liens, méritent également des recherches plus poussées (voir Thompson [1997]).

3.1 Plan à vague unique

Dans un plan à dépistage de liens à vague unique, un échantillon initial de nœuds est tiré à l'aide d'un plan négligable quelconque de la population de nœuds du graphe. Pour chaque nœud de l'échantillon, les nœuds qui lui sont contigus sont ajoutés à l'échantillon. On suppose que la procédure en boule de neige s'arrête après une vague. Ainsi, le nœud v est ajouté si pour un nœud u quelconque de l'échantillon initial $x^{uv} = 1$.

Soit s_0 , l'ensemble de nœuds de l'échantillon initial et s_1 , les nœuds ajoutés qui ne se trouvent pas dans l'échantillon initial. L'échantillon global est $s = s_0 \cup s_1$. L'ensemble d'étiquettes au complet peut s'exprimer comme l'union de trois ensembles disjoints, $U = s_0 \cup s_1 \cup s_2$. Les valeurs y associées aux nœuds peuvent être ordonnées en conséquence sous forme de séquence $(y_{s_0}, y_{s_1}, y_{s_2})$, où $y_a = (y_u : u \in a)$ est la sous-séquence de y limitée aux indices du sous-ensemble $a \subset U$. La matrice de contiguïté x est ordonnée en conséquence et partitionnée en sous-matrices $x_{s_0 s_0}, x_{s_0 s_1}, x_{s_0 s_2}, x_{s_1 s_0}, x_{s_1 s_1}, x_{s_1 s_2}$, et ainsi de suite, où $x^{ab} = (x_{uv} : u \in a, v \in b)$. Le fait d'ordonner la matrice de contiguïté de cette façon facilite la description des facteurs de la vraisemblance.

Pour le plan ci-dessus, la probabilité de sélection de l'échantillon s dépend uniquement de $x_{s_0 U}$ et, par conséquent, on peut écrire $p(s | x_{s_0 U})$, où $x_{s_0 U}$ peut aussi être remplacé par sa permutation de colonne $(x_{s_0 s_0}, x_{s_0 s_1}, x_{s_0 s_2})$. Autrement dit, la probabilité de sélection de l'échantillon final $s = s_0 \cup s_1$ dépend de liens entre l'échantillon initial et d'autres unités du graphe, tant pour s que pour s_0 . Les données comprennent $(s, y_s, x_{s_0 U})$. Puisque le plan ne dépend d'aucune valeur x ou y en dehors des données, ni de valeurs paramétriques de modèle, le plan de sondage est négligable pour ce qui est de l'inférence fondée sur la vraisemblance.

vraisemblance et de Bayes, est simplifiée si le plan peut être laissé de côté à l'étape de l'inférence. Le fait que le plan d'échantillonnage n'influence pas la valeur d'un estimateur de Bayes ou fondé sur la vraisemblance dans un sondage a été observé par Godambe (1966) pour des plans qui ne dépendent d'aucune valeur de la variable d'intérêt et par Basu (1969) pour des plans qui ne dépendent d'aucune valeur de la variable d'intérêt à l'extérieur de l'échantillon. Scott et Smith (1973) ont montré que le plan pouvait avoir de la pertinence pour l'inférence lorsque les données ne comportaient pas suffisamment d'information sur les étiquettes des unités de l'échantillon. Rubin (1976) a décrit les conditions exactes dans lesquelles un mécanisme de données manquantes (dont un plan d'échantillonnage pourrait être considéré comme un exemple) pourrait avoir de la pertinence fondée sur la vraisemblance. Pour des méthodes fondées sur les fréquences. Pour des méthodes fondées sur la vraisemblance comme les méthodes du maximum de vraisemblance et de Bayes, le plan est «négligeable» si le plan ou le mécanisme ne dépend pas de valeurs de la variable d'intérêt à l'extérieur de l'échantillon ou de paramètres quelconques de la distribution de ces valeurs. Pour une inférence axée sur les fréquences comme une estimation non biaisée pour ce qui est du plan ou du modèle, cependant, le plan est pertinent s'il dépend de valeurs quelconques de la variable d'intérêt, même dans l'échantillon. Scott (1977) a montré que le plan est pertinent pour une estimation de Bayes si l'information auxiliaire utilisée dans le plan n'est pas accessible à l'étape de l'inférence. Sargent et Smith (1984) ont indiqué de façon générale et détaillée à quel moment le plan est pertinent dans le cas de sondages. Thompson et Seber (1996) ont décrit des plans adaptables dans lesquels la procédure de sélection utilise délibérément des valeurs observées de la variable d'intérêt, et ils ont discuté de la pertinence du plan pour l'inférence à partir de différentes perspectives fondées sur le plan et sur un modèle. Des questions semblables de plan et d'inférence se posent pour des plans expérimentaux adaptables, par exemple les expériences médicales dans lesquelles des considérations d'éthique entraînent un choix de traitements adaptable afin de favoriser un traitement plus prometteur à mesure que l'étude avance (voir Floumoy et Rosenberger 1995, Rosenberger 1996, Wei et coll. 1990). Il importe de ne pas oublier qu'un plan qui est «négligeable» pour une inférence fondée sur la vraisemblance n'est pas nécessairement négligeable pour une inférence axée sur les fréquences, par exemple une estimation non biaisée pour ce qui est du modèle, et que même si un plan est négligeable à l'étape de l'inférence, en ce sens que la façon dont l'estimateur est calculé, par exemple, ne dépend pas du plan utilisé, ce même plan demeure pertinent a priori pour ce qui est des propriétés de l'estimateur.

Les données d'échantillon $d = (s, y_s, x_s)$ sont fonction de l'échantillon sélectionné et des valeurs de graphe y et x . La vraisemblance peut s'écrire

$$L(\psi, d) = \sum p(s|y, x; \psi) f(y, x; \psi) \tag{1}$$

où la somme englobe les résultats (y, x) conformes aux données d . Puisque les valeurs y et x pour les nœuds et les paires de nœuds dans l'échantillon sont régies par les données, la somme englobe toutes les valeurs possibles des variables non observées y_s et x_s et, en réalité, elle représente la probabilité marginale de l'échantillon s sélectionné et les variables observées associées y_s et x_s . Ainsi, en général, la fonction de vraisemblance dépend aussi bien du plan que du modèle. La quantité $\sum_{y_s, x_s} f(y, x; \psi)$, fondée sur le modèle seulement sans tenir compte du plan, a été appelée «la vraisemblance à valeur nominale» par Dawid et Dickey (1977) car une inférence fondée sur cette seule fonction accueille les données à leur valeur nominale sans tenir compte de leur mode de sélection.

Pour tout plan dans lequel le tirage de l'échantillon dépend des valeurs de graphe y et x seulement par l'entremise des valeurs y_s et x_s comprises dans les données, la probabilité du plan peut être retirée de la somme et constituer un facteur distinct dans la vraisemblance. Si, de plus, les paramètres de plan et de modèle sont distincts et sans relation, la vraisemblance peut s'écrire

$$L(\phi, \psi, d) = p(s|y_s, x_s; \phi) \sum_{y_s, x_s} f(y_s, x_s; \psi) \tag{2}$$

où ϕ désigne les paramètres de plan et ψ désigne les paramètres de modèle. Dès lors, le plan n'influence pas la valeur des estimateurs ou prédicteurs fondés sur des méthodes de vraisemblance directe comme c'est le cas des estimateurs du maximum de vraisemblance ou de Bayes. Pour tout plan «négligeable» de ce genre, la somme dans la vraisemblance ci-dessus, pour l'ensemble des valeurs de y et x entraînant la valeur des données en question, est simplement la probabilité marginale des valeurs y et x associées aux données de l'échantillon. Cette distribution marginale dépend de l'échantillon qui a été sélectionné, mais elle ne dépend pas du mode de sélection de l'échantillon. Pour ce qui est de l'inférence fondée sur la vraisemblance et comportant un plan qui est négligeable en ce sens, la vraisemblance à valeur nominale donne lieu à l'inférence correcte.

3. QUELQUES PLANS À DÉPISTAGE DE LIENS

La présente section décrit un choix de plans à dépistage de liens. Chacun de ces plans est négligeable pour ce qui est de la vraisemblance à la condition que l'échantillon initial soit tiré à l'aide d'une procédure négligeable, et que les données englobent toutes les valeurs mises en jeu dans la procédure de sélection. Puisque, pour tous les plans décrits dans la présente section, l'échantillon de paires de nœuds

couramment ou à des sommes de données comme des moyennes d'échantillons et des proportions de nœuds ou de valeurs de lien, on se rend compte que, dans la plupart des cas, les estimations conventionnelles ne sont pas les meilleures. De même, il se peut que des estimateurs qui seraient appropriés si les données comprenaient le graphe tout entier ne soient pas appropriés si les données ne représentent qu'un échantillon du graphe. Ces résultats signifient que des estimations conventionnelles ou des sommes non corrigées de données d'échantillon obtenues à l'aide de procédures de dépistage de liens risquent d'être trompeuses s'ils sont considérés dans le cadre de la population ou des caractéristiques du graphe tout entier. Les interprétations de cette divergence fournies par les auteurs du présent exposé permettent de mieux comprendre les conditions dans lesquelles la meilleure estimation aurait tendance à être inférieure ou supérieure à l'estimation conventionnelle.

On trouvera à la section 2 la notation et les questions de base pour le plan d'échantillonnage et l'inférence dans un contexte de graphes. À la section 3, les auteurs décrivent tout un choix de procédures de dépistage de liens, que l'on peut toutes analyser à l'aide de la stratégie décrite dans le présent exposé. À la section 4, on trouvera une description d'une catégorie de modèles de graphe que les auteurs utilisent pour illustrer les méthodes d'inférence de l'exposé. La section 5 traite des méthodes d'estimation et de prévision du maximum de vraisemblance pour des paramètres de graphe et des valeurs de population réalisées.

2. MODÈLES DE GRAPHE ET PLANS D'ÉCHANTILLONNAGE

Nous considérons un graphe de N nœuds (unités) étiquetés $1, 2, \dots, N$. Une variable d'intérêt Y_u est associée au u -ième nœud. L'ensemble complet d'étiquettes de nœud est noté $U = \{1, 2, \dots, N\}$ et la séquence de variables de nœud $\mathbf{Y} = (Y_1, \dots, Y_N)$. Pour deux nœuds distincts u et v , la variable indicatrice X_{uv} est égale à 1 s'il y a un arc (lien directionnel) entre u et v et à 0 autrement. La matrice des indicateurs d'arc comportant X_{uv} comme élément de la u -ième ligne et de la v -ième colonne est la matrice de contiguité du graphe, notée \mathbf{X} . Par souci de simplicité, nous supposons que les éléments diagonaux X_{uu} sont de valeur zéro. On parle parfois de la paire ordonnée (u, v) comme d'une dyade de type $(X_u^n, X_v^n; X_{uv}^{nn}, X_{vu}^{nn})$. Un modèle de graphe est donné par une densité ou probabilité conjointe $f(\mathbf{y}, \mathbf{x}; \psi)$ pour les résultats \mathbf{y} et \mathbf{x} de \mathbf{Y} et \mathbf{X} , respectivement, et il peut dépendre d'un ou de plusieurs paramètres ψ inconnus.

Un échantillon s du graphe est un sous-ensemble de nœuds et un sous-ensemble de paires de nœuds. Nous pouvons exprimer l'échantillon combiné sous la forme $s = (s_{(1)}, s_{(2)})$, où $s_{(1)}$ désigne le sous-ensemble de nœuds sélectionné pour l'observation des valeurs \mathbf{y} associées, et $s_{(2)}$ désigne le sous-ensemble de paires de nœuds

sélectionné pour l'observation des valeurs \mathbf{x} associées. Les données comprennent les étiquettes de nœuds et de paires de nœuds figurant dans l'échantillon combiné de même que les valeurs associées de nœuds et d'indicateurs d'arc, c'est-à-dire $d = (u, (v, w), Y_u^{nn}, X_{vw}^{nn}; n \in s_{(1)}, (v, w) \in s_{(2)})$ ou encore, plus simplement, $d = (s, Y_s^{(n)}, X_s^{(w)})$. De plus, il est souvent commode d'utiliser Y_s pour désigner les valeurs \mathbf{y} des nœuds dans l'échantillon combiné et \mathbf{x}_s pour les valeurs \mathbf{x} des paires de nœuds dans l'échantillon combiné. $Y_s^{(s)}$ et $\mathbf{x}_s^{(s)}$ désignant les valeurs des nœuds et des paires de nœuds pour lesquelles des valeurs \mathbf{x} sont enregistrées et les nœuds pour lesquels des valeurs \mathbf{y} sont enregistrées et enregistrées peuvent être des ensembles sans guère de relation. En particulier, les procédures de dépistage de liens considérées dans le présent exposé entraînent souvent des données sur les liens entre des nœuds pour $s_{(1)}$ et des nœuds à l'extérieur de $s_{(1)}$.

Le plan d'échantillonnage est la procédure qui permet de tirer un échantillon. Cette procédure peut être contrôlée par les chercheurs, comme c'est le cas d'un plan probabiliste mis en œuvre délibérément, ou elle peut échapper au contrôle des chercheurs et être déterminée par les circonstances. Si la probabilité que l'échantillon soit tiré ne dépend pas des valeurs de nœud \mathbf{y} ou des valeurs de lien \mathbf{x} ou encore des paramètres ψ relevant du modèle de graphe, on parle de plan «conventionnel». Pour un tel plan la probabilité que l'échantillon s soit tiré s'écrit $p(s)$ ou $p(s; \phi)$, où ϕ désigne tout paramètre inconnu relevant du plan (mais non pas du modèle), comme c'est le cas d'un échantillonnage de Bernoulli à probabilité d'inclusion inconnue ϕ pour chaque nœud. Le plan peut dépendre d'une ou de plusieurs variables auxiliaires qui sont connues pour l'ensemble de la population, mais cette dépendance reste implicite dans la notation $p(s)$. Les plans conventionnels englobent les plans probabilistes classiques comme l'échantillonnage aléatoire simple, systématique, stratifié, à plusieurs degrés et à probabilités inégales, de même que des plans par choix raisonné et équilibrés, fondés sur un modèle, qui sont axés sur des variables auxiliaires.

Si la probabilité qu'un échantillon soit tiré dépend de valeurs de nœud et de lien dans la population. De plus, la procédure de tirage s'adapte à la configuration réalisée des valeurs de nœud et de lien dans la population. Ainsi, en général, le plan d'échantillonnage dans le contexte d'un graphe comporte une probabilité de tirage qui s'écrit $p(s | \mathbf{y}, \mathbf{x}; \psi)$, où \mathbf{y} désigne la séquence des valeurs de nœud, \mathbf{x} désigne la matrice des valeurs d'arc et ψ désigne tout paramètre mis en jeu.

Le mode d'inférence fondé sur la vraisemblance, comme les méthodes de prévision ou d'estimation du maximum de

d'intérêt observée pour le nœud. Ainsi, dans une étude épidémiologique d'une maladie transmise sexuellement, les liens sociaux ou sexuels peuvent être suivis uniquement pour des répondants infectés. Des méthodes d'estimation non biaisées pour ce qui est du plan ont été décrites pour tout un choix de stratégies d'échantillonnage en grappes adaptées.

Les méthodes d'inférence fondées sur le plan de sondage comme les techniques d'estimation à base de plan pour l'échantillonnage en réseau, l'échantillonnage en boule de neige et l'échantillonnage en grappes adaptable offrent l'avantage que la validité de propriétés comme l'absence de biais dû au plan de sondage ou la convergence ne dépendent pas d'un modèle prévu quelconque de la population. Par contre, ces propriétés dépendent bel et bien de l'exécution prévue du plan d'échantillonnage. Quant aux méthodes fondées sur un modèle, décrites dans le présent exposé, elles dépendent effectivement d'un modèle prévu pour la population ou le graphique. Leur avantage pratique relève de ce qu'elles s'appliquent à tout un choix de procédures de sélection de l'échantillon, d'où leur plus grande souplesse quant au tirage concret de l'échantillon.

En réalité, de nombreuses études réelles de populations cachées et difficiles d'accès font appel à des procédures de tirage d'un échantillon, y compris le dépistage de liens, qu'il n'est pas aisé d'analyser en fonction de probabilités idéalisées induites à l'aide du plan. Ainsi, dans une étude du rapport entre une structure en réseau et des comportements à risque comme le partage d'aiguilles parmi les utilisateurs de drogues injectables dans le quartier Bushwick de Brooklyn, des répondants «indices» (initiaux) ont servi de «techniciens auxiliaires» pour intégrer des membres de leur réseau à l'étude (Friedman, Neagus, Jose, Curtis, Goldstein, Friedman, Rothenberg et Des Jarlais 1997, Neagus, Friedman, Goldstein, Curtis, Goldstein, Iddefonso and Des Jarlais 1996). Toutefois, environ 61 % seulement des individus liés ont été effectivement recrutés. Dans une étude à long terme de la transmission hétérosexuelle de l'infection par VIH (Rothenberg, Woodhouse, Portrat, Muth, Darrow et Kivodah 1995), la population cible à l'étude comprenait des travailleurs de l'industrie du sexe, leurs partenaires payants et non payants, des personnes utilisant des drogues injectables et les partenaires sexuels des utilisateurs de drogues dans la région de Colorado Springs. On a également sélectionné et, en plus de leurs caractéristiques individuelles, des personnes figurant dans l'échantillon initial sciemment trouvées et interviewées des personnes nommées par deux répondants ou plus. L'éventail de procédures de dépistage des liens utilisées dans des études de ce genre explique l'insistance des auteurs du présent exposé sur des méthodes d'inférence fondées sur un modèle.

Lorsque l'on compare les prédicteurs et les estimateurs du maximum de vraisemblance obtenus dans le présent exposé à des estimations conventionnelles utilisées

leur tour sont prises d'identifier le même nombre de utilisées pour l'échantillon initial. Des plans en boule de neige ont été élaborés dans un contexte de graphique à l'aide de différents plans d'échantillonnage probabilistes initiaux et d'un nombre quelconque de liens et de vagues par Frank (1971, 1977a,b, 1978a,b, 1979a), qui a pu obtenir un choix de méthodes fondées sur le plan et sur un modèle en vue de l'estimation. Snijders (1992) a inclus dans la notion d'échantillonnage en boule de neige les plans dans lesquels on suit un sous-échantillon seulement de liens pour chaque nœud. Kivodah (1989) a utilisé la notion de marche aléatoire pour désigner le cas dans lequel un seul des liens d'un nœud est sélectionné au hasard et suivi jusqu'à un autre nœud, dont un des liens est sélectionné à son tour, et ainsi de suite. Les méthodes d'échantillonnage à dépistage de liens dans les-

quelles il n'y a qu'un seul lien pour chaque nœud ont été décrites en termes de chaînes (Erickson 1979). Frank et Snijders (1994) ont examiné l'estimation fondée sur un modèle et sur le plan pour une taille de population cachée, c'est-à-dire le nombre de nœuds dans le graphique, d'après des échantillons en boule de neige. D'autres enjeux statistiques et pratiques de l'échantillonnage à partir de réseaux sociaux en fonction de différents types de plans à dépistage de liens en boule de neige, dirigés en chaîne et autres sont décrits dans Granovetter (1976), Morgan et Rytina (1977), Frank (1979b, 1981, 1988), Walters et Biernacki (1989), van Meeter (1990), Spreen (1992), Wasserman et Faust (1994), Spreen (1998) et Robins (1998).

Des méthodes d'estimation fondées sur le plan ont été mises au point également pour les plans très connexes de l'échantillonnage en réseau au multiple, dans lequel on suit des liens sociaux et administratifs et des liens de parenté (Birnbaum et Sirken 1965, Kalton et Anderson 1986, Levy 1977, Levy et Lemeshow 1991, Sirken 1970, 1972a, b, Sirken et Levy 1974, Sudman, Sirken et Cowan 1988). Ainsi, pour une enquête sur une maladie rare, un échantillon initial de ménages pourrait être sélectionné au hasard, des données étant obtenues tant pour les résidents du ménage que pour leurs frères et sœurs. L'estimation fondée sur le plan dans ce genre de stratégie est facilitée par la symétrie des liens et l'incorporation de composantes entièrement liées dans l'échantillon, et l'on a obtenu des estimateurs non biaisés pour un échantillonnage en réseau à l'aide de plusieurs plans initiaux différents.

Une autre procédure de dépistage des liens pour laquelle il existe des estimateurs fondés sur le plan est l'échantillonnage en grappes adaptable (Thompson 1990, 1997, Thompson et Seber 1996), que l'on a formulé dans un contexte de graphique aussi bien que d'espace. Lorsqu'un échantillon initial de nœuds a été sélectionné à l'aide de l'un ou l'autre plan initial, la décision de suivre les liens ou non à partir d'un nœud dépend de la valeur d'une variable

Estimation fondée sur un modèle et comportant des plans d'échantillonnage à dépistage de liens

STEVEN K. THOMPSON et OVE FRANK¹

RÉSUMÉ

Il arrive souvent que l'on obtienne des échantillons de populations humaines cachées et difficiles d'accès à l'aide de procédures permettant de suivre des liens sociaux d'un répondant à un autre. Une inférence de l'échantillon à la population d'intérêt elle-même risque d'être influencée par le type de plan à dépistage de liens et le type de données qui en résulte. La répartition mixte de valeurs de nœud représentant des caractéristiques des individus et des indicateurs d'arc correspondant aux relations sociales entre les individus. Les auteurs décrivent des estimateurs de quantités réalisées de graphes démographiques à paramètres de graphes démographiques. On obtient des prédicteurs de quantités réalisées de graphes démographiques à l'aide de la vraisemblance prédictive. Ces estimateurs et prédicteurs sont comparés à des sommaires de données traditionnels et illustrés à l'aide d'un exemple numérique.

MOTS CLÉS : Échantillonnage en boule de neige; échantillonnage adaptable; échantillonnage par graphe; plans d'échantillonnage négligibles; plans d'échantillonnage à dépistage de liens; échantillonnage en réseau; vraisemblance; vraisemblance prédictive.

1. INTRODUCTION

Dans l'étude de populations humaines cachées et difficiles d'accès, il arrive couramment que l'échantillon soit obtenu à l'aide de procédures de dépistage de liens, des liens sociaux étant suivis d'un répondant à un autre. Ainsi, dans une étude de l'usage de drogues injectables relativement à la propagation de l'infection par VIH, il est possible que l'on demande aux répondants initiaux d'identifier des partenaires sexuels ou d'injecter que l'on ajoute à l'échantillon. Dans une telle étude, les liens sociaux ont une importance inhérente si l'on veut comprendre les phénomènes étudiés, tout en étant utiles ou essentiels à la sélection de l'échantillon. Toutefois, une inférence de l'échantillon à la population ou à la composition sociale elle-même risque d'être influencée par les procédures de dépistage de liens et par le type de données qui en résulte. Les auteurs évaluent ce problème d'inférence en fonction du plan d'échantillonnage, et décrivent certaines méthodes d'inférence pour ce genre d'étude fondée sur l'estimation et la prévision du maximum de vraisemblance.

Les populations humaines à composition sociale sont souvent modélisées sous forme de graphe, les nœuds de celui-ci représentant des individus et les contours ou les arcs du graphe représentant des liens sociaux, des relations ou des transactions. Le graphe démographique comme tel peut être considéré soit comme une structure fixe, soit comme une réalisation d'un modèle de graphe stochastique. Dans des études de populations humaines, notamment celles qui sont cachées ou difficiles d'accès, il est

rarement possible d'obtenir des données sur l'ensemble de la population ou de la composition du graphe. On obtient plutôt les données d'un échantillon, qui a pu être tiré à l'aide de méthodes novatrices et peu conventionnelles, y compris celles qui sont axées sur des arcs ou des liens établis d'un individu à un autre. Il est possible que les données des individus de l'échantillon, des liens sociaux au sein de l'échantillon et, dans certains cas, des liens entre les individus de l'échantillon et d'autres individus.

Dans le présent exposé, le terme « plan d'échantillonnage » désigne la procédure en vertu de laquelle l'échantillon est tiré, qu'elle soit voulue ou fortuite. Dans de nombreuses études ethnographiques et sociologiques de populations cachées, les plans d'échantillonnage à dépistage de liens sont considérés comme la seule façon pratique d'obtenir un échantillon suffisamment grand. Dans d'autres études, la composition sociale elle-même représente l'objet de l'étude, et les méthodes de dépistage de liens servent à obtenir des échantillons convenablement structurés.

La documentation statistique traitant des plans d'échantillonnage et de l'estimation à l'aide de plans d'échantillonnage à dépistage de liens fait état de procédures comme l'échantillonnage en boule de neige, l'échantillonnage dirigé en chaîne, les marches aléatoires, l'échantillonnage par liens, l'échantillonnage en réseau ou multiple et l'échantillonnage adaptable. Goodman (1961) a proposé la notion d'échantillonnage en boule de neige pour un type de plan à dépistage de liens dans lequel les individus d'un échantillon initial sont

¹ Steven K. Thompson, Department of Statistics, 326 Thomas Building, Pennsylvania State University, University Park, PA 16802 USA; Ove Frank, Department of Statistics, Stockholm University, S-10691 Stockholm, Sweden. Cette recherche est le résultat d'un effort de collaboration continu. L'ordre dans lequel le nom des auteurs paraît a été déterminé par un pile ou face.

REMERCIEMENT

L'auteur remercie les évaluateurs de leurs commentaires et suggestions. La première ébauche du présent article a été achevée au U.S. Census Bureau et au U.S. Bureau of Labor Statistics à l'époque où l'auteur y travaillait à titre de chercheur supérieur titulaire d'une bourse de l'ASA/NSF. Les travaux ont également été financés par les bourses DMS-9504425 et DMS-9803112 de la National Science Foundation et par la bourse MDA904-99-1-0032 de la National Security Agency.

ANNEXE

1. **Preuve de (7):** Quand $n/N \approx 0$, $V(I_R^x - Y) \approx V(I_R^x)$. Alors (7) découle de

$$V(I_R^x) = \frac{n^2}{N^2} \left\{ \sigma^2 E \left[\left(\sum_{i \in S} x_i \right)^2 / \left(\sum_{i \in S} x_i \right) \right] + \beta^2 V \left(\sum_{i \in S} x_i \right) \right\} \approx \frac{n}{N^2} \left(\frac{d}{\sigma^2 \mu_x} + \beta^2 v_x \right)$$

pour une valeur élevée de n , où la dernière égalité approximative découle du fait que, subordonnée aux x_i , $E(\sum_{i \in S} x_i) = d \sum_{i \in S} x_i$.

2. **Preuve de (9):** Étant donné le modèle (1),

$$V(I^C) = \frac{n^2}{N^2} \left\{ V \left(\sum_{i \in S} x_i^{1/2} e_i \right) + V \left(\beta \sum_{i \in S} x_i + \sum_{i \in S^c} x_i \right) \right\} = \frac{n^2}{N^2} \left\{ \sigma^2 \mu_x + \beta^2 V \left(\sum_{i \in S} x_i \right) + V \left(\sum_{i \in S^c} x_i \right) \right\} + 2\beta \text{Cov} \left(\sum_{i \in S} x_i, \sum_{i \in S^c} x_i \right)$$

$$= \frac{n}{N^2} \left\{ \sigma^2 d \mu_x + \beta^2 [d v_x + (1 - d) \mu_x^2] - 2\beta d (1 - d) \mu_x^2 \right\} + (\beta - 1)^2 d (1 - d) \mu_x^2.$$

BIBLIOGRAPHIE

- BUTANI, S., HARTER, R., et WOLTER, K. (1998). Estimation procedures for the Bureau of Labor Statistics current employment statistics program. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.
- COCHRAN, W. G. (1977). *Sampling Techniques*. Troisième édition. New York: Wiley.
- KALTON, G., et KASPRZYK, D. (1986). Le traitement des données d'enquête manquantes. *Techniques d'enquête*, 12, 1-17.
- KING, C., et KORNBAB, M. (1994). Inventory Of Economic Area Statistical Practices. ESMR Report Series 9401, Bureau of the Census, Washington D.C.
- KREWSKI, D., et RAO, J. N. K. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.
- LEE, H., RANCOURT, E., et SÄRNDA, C.-E. (1994). Experiments with variance estimation from survey data with imputed values. *Journal of Official Statistics*, 10, 231-243.
- RAO, J. N. K., et SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.
- RUBIN, D. B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- VALLIANT, R. (1993). Poststratification and conditional variance estimation. *Journal of American Statistical Association*, 88, 89-96.
- Shao: Imputation par la méthode cold deck et la méthode du quotient
3. **Preuve de (11):** Étant donné les conditions hypothétiques auxquelles sont assujetties (y_i, x_i) et (y_i^*, x_i^*) ,
- $$\text{eqm}(I^{C-R}) = \frac{n^2}{N^2} V \left(\sum_{i \in S} y_i + \sum_{i \in S^c} z_i \right) = \frac{n^2}{N^2} V \left(\sum_{i \in S} z_i + \sum_{i \in S} z_i \right) = \frac{n^2}{N^2} V \left(\sum_{i \in S} z_i \right) + \frac{n^2}{N^2} V \left(\sum_{i \in S} z_i \right) + 2\text{Cov} \left(\sum_{i \in S} z_i, \sum_{i \in S} z_i \right)$$
- $$= \frac{n}{N^2} \left\{ \sigma^2 d (\mu_x + \gamma_x) + (\beta^2 v_x + \sigma^2 \gamma_x) - 2\sigma^2 d \gamma_x \right\} = \frac{n}{N^2} \left\{ \sigma^2 d \mu_x + \beta^2 v_x + \sigma^2 (1 - d) \gamma_x \right\}.$$

qui donne les meilleurs résultats. Cependant, si x correspond à la liste annuelle de paie de l'année courante (tableau 3), les estimations produites par la méthode cold deck simple sont manifestement trop faibles; d'après l'EQM estimative, la méthode cold deck simple est celle qui donne les moins bons résultats, parce qu'elle est entachée d'un biais trop grand (présenté au tableau 3).

Tableau 2
Totaux et EQM estimatifs quand x = revenu annuel de l'année précédente, y = liste de paie de l'année courante et z = liste de paie annuelle de l'année précédente

Branche	d'activité	Méthode Cold Deck- Méthode par		Méthode	
		Estimation	Cold Deck	par quotient	quotient
1	Total	5,31	$\times 10^9$	5,19	$\times 10^9$
		v_1	$7,73 \times 10^{14}$	8,46	$\times 10^{14}$
		v_2	$1,39 \times 10^{15}$	2,50	$\times 10^{15}$
		EQM proposée	$2,30 \times 10^{15}$	3,34	$\times 10^{15}$
		EQM naïve	$7,73 \times 10^{14}$	8,46	$\times 10^{14}$
		Rapport des	$1,79$	$3,95$	$2,97$
2	Total	1,66	$\times 10^{10}$	1,63	$\times 10^{10}$
		v_1	$4,00 \times 10^{15}$	4,19	$\times 10^{15}$
		v_2	$6,03 \times 10^{15}$	2,88	$\times 10^{15}$
		EQM proposée	$1,02 \times 10^{16}$	3,30	$\times 10^{16}$
		EQM naïve	$4,00 \times 10^{15}$	4,19	$\times 10^{15}$
		Rapport des	$2,54$	$7,87$	$1,12$
3	Total	3,54	$\times 10^{10}$	3,53	$\times 10^{10}$
		v_1	$1,32 \times 10^{16}$	1,80	$\times 10^{16}$
		v_2	$5,44 \times 10^{16}$	6,82	$\times 10^{16}$
		EQM proposée	$6,97 \times 10^{16}$	1,04	$\times 10^{17}$
		EQM naïve	$1,32 \times 10^{16}$	1,80	$\times 10^{16}$
		Rapport des	$5,27$	$5,80$	$1,40$
4	Total	1,27	$\times 10^{10}$	1,22	$\times 10^{10}$
		v_1	$2,11 \times 10^{16}$	2,14	$\times 10^{16}$
		v_2	$3,91 \times 10^{15}$	8,26	$\times 10^{15}$
		EQM proposée	$2,59 \times 10^{16}$	2,97	$\times 10^{16}$
		EQM naïve	$2,11 \times 10^{16}$	2,14	$\times 10^{16}$
		Rapport des	$1,23$	$1,39$	$2,01$

2. Aucune conclusion catégorique ne peut être tirée en ce qui concerne les résultats relatifs (en ce qui concerne l'EQM estimative) de la méthode d'imputation par quotient et de la méthode d'imputation cold deck par quotient. Dans l'exemple choisi ici, la méthode cold deck par quotient donne de meilleurs résultats pour les branches d'activité 1 à 3, tandis que la méthode d'imputation par quotient donne de meilleurs résultats pour la branche d'activité 4. Certains diagrammes de dispersion des données (non présentés) indiquent que la corrélation entre x et z est plus forte pour les branches d'activité 1 à 3 que pour la branche d'activité 4, situation qui pourrait expliquer l'écart entre les résultats relatifs des deux méthodes d'imputation. Voir aussi la discussion qui suit la formule (12).

Branche	d'activité	Méthode Cold Deck- Méthode par		Méthode	
		Estimation	Cold Deck	par quotient	quotient
1	Total	4,49	$\times 10^9$	5,19	$\times 10^9$
		v_1	$8,10 \times 10^{14}$	8,46	$\times 10^{14}$
		v_2	$1,38 \times 10^{15}$	2,64	$\times 10^{15}$
		EQM proposée	$1,03 \times 10^{16}$	3,49	$\times 10^{16}$
		EQM naïve	$8,10 \times 10^{14}$	8,46	$\times 10^{14}$
		Rapport des	$12,68$	$4,12$	$1,81$
2	Total	1,59	$\times 10^{10}$	1,63	$\times 10^{10}$
		v_1	$4,36 \times 10^{15}$	4,19	$\times 10^{15}$
		v_2	$8,20 \times 10^{15}$	1,48	$\times 10^{16}$
		EQM proposée	$2,73 \times 10^{16}$	1,90	$\times 10^{16}$
		EQM naïve	$4,36 \times 10^{15}$	4,19	$\times 10^{15}$
		Rapport des	$6,25$	$4,54$	$1,18$
3	Total	3,10	$\times 10^{10}$	3,53	$\times 10^{10}$
		v_1	$1,25 \times 10^{16}$	1,80	$\times 10^{16}$
		v_2	$4,56 \times 10^{16}$	9,25	$\times 10^{16}$
		EQM proposée	$1,89 \times 10^{17}$	1,10	$\times 10^{17}$
		EQM naïve	$1,25 \times 10^{16}$	1,80	$\times 10^{16}$
		Rapport des	$15,13$	$6,15$	$1,56$
4	Total	1,06	$\times 10^{10}$	1,22	$\times 10^{10}$
		v_1	$1,93 \times 10^{16}$	2,14	$\times 10^{16}$
		v_2	$2,67 \times 10^{15}$	4,62	$\times 10^{15}$
		EQM proposée	$4,03 \times 10^{16}$	2,60	$\times 10^{16}$
		EQM naïve	$1,93 \times 10^{16}$	2,14	$\times 10^{16}$
		Rapport des	$2,09$	$1,22$	$1,72$

Totaux et EQM estimatifs si x = liste annuelle de paie de l'année courante, y = revenu annuel de l'année précédente et z = liste annuelle de paie de l'année précédente

Tableau 3

3. La valeur de l'EQM estimative naïve est beaucoup plus faible que celle de l'EQM estimative proposée et trop optimiste. Par exemple, au tableau 3, les valeurs de l'EQM naïve pour la méthode cold deck simple sont systématiquement plus faibles que celles obtenues pour la méthode cold deck par quotient, alors que l'on sait que la méthode cold deck simple ne donne pas d'aussi bons résultats dans ce cas-là. Dans l'exemple choisi ici, v_2/v_1 n'est pas faible, parce que certaines fractions d'échantillonnage sont importantes. Puisque l'EQM estimative naïve est égale à v_1 (dans le cas des méthodes d'imputation cold deck) ou diffère peu de v_1 (pour la méthode d'imputation par quotient), la sous-estimation due à l'utilisation de l'EQM estimative naïve tient principalement au fait de traiter les valeurs imputées comme des valeurs observées dans les strates où la fraction d'échantillonnage est importante (et en ne tenant pas compte du biais qui entache les estimateurs par la méthode cold deck simple dans le cas du tableau 3).

formule (14) et que l'on traite les valeurs imputées pour la non-réponse comme des données observées, on obtient l'estimateur de $\text{eqm}(x_R)$:

$$\hat{v}_{1R} = \sum_{i=1}^h \left(1 - \frac{N_h}{n_h} \right) \frac{n_h - 1}{n_h} \sum_{j=1}^{I_{h(s)}} w_{1j}' - \frac{1}{n_h} \sum_{j=1}^{I_{h(s)}} w_{1j}' \quad (18)$$

où $z_i' = a_1 y_i' + (1 - a_1) \hat{\beta}_1 x_i'$, expression qui est différente du premier terme v_{1R} de notre estimateur $\text{eqm}(x_R)$ et, donc, n'est pas asymptotiquement valide, même si n/N est négligeable.

4. EXEMPLE

Nous nous servons d'un ensemble de données provenant de l'enquête annuelle sur les transports (EAT) réalisée par le U.S. Census Bureau pour illustrer nos calculs.

L'EAT est une enquête réalisée auprès d'entreprises comprenant un ou plusieurs établissements dont l'activité principale consiste à fournir des services de transport commercial motorisé de marchandises ou des services d'entreposage public aux Etats-Unis. Un échantillon aléatoire simple stratifié est sélectionné sans remise parmi les employeurs qui figurent sur la liste statistique type des établissements du Census Bureau. Les strates, qui correspondent aussi aux catégories d'imputation dans le présent exemple, sont définies d'après la taille des entreprises dans chaque branche d'activité.

L'enquête comporte diverses variables. Nous considérons l'estimation des totaux de population des revenus annuels de l'année courante (y) dans quatre branches de l'année courante comme covariable x pour l'imputation par la méthode cold deck simple et par la méthode du quotient. Nous nous servons de la liste de paie annuelle pour l'année courante et de la liste de paie annuelle pour l'année précédente à titre de \bar{y} et \bar{x} , respectivement. Pour quatre branches d'activité et trois méthodes d'imputation, le tableau 2 énumère les totaux estimatifs, les EQM estimatives proposées pour les totaux estimatifs, les EQM estimées par la méthode naïve pour les totaux estimatifs (obtenus en traitant les valeurs imputées comme des données observées) et les ratios des EQM (EQM estimative proposée sur EQM estimative naïve). Notons que l'EQM estimative proposée est égale à la somme de v_1 et v_2 pour les méthodes d'imputation par quotient et d'imputation cold

deck par quotient ou à la somme de v_1 , v_2 et du carré du biais estimatif dans le cas de la méthode cold deck simple. Les valeurs de v_1 et v_2 figurent également dans le tableau.

Tableau 1

Effet de l'échantillon, taille de la réponse et poids d'échantillon selon la branche d'activité et la strate

Branches d'activité	Strate	Effet de l'échantillon	Taille de la réponse	Poids d'échantillon
1	0	31	24	1.00
	1	14	6	12.43
	2	11	7	8.91
	3	10	4	6.10
	4	11	6	5.73
	5	16	12	2.70
	6	18	13	2.17
2	0	86	82	1.00
	1	8	2	32.91
	2	13	10	9.85
	3	11	9	10.82
	4	12	10	6.08
	5	13	10	3.60
3	0	38	30	1.00
	1	14	9	87.91
	2	11	8	67.39
	3	13	10	44.48
	4	14	13	25.28
	5	16	13	15.57
	6	18	12	9.80
	7	15	11	6.23
	8	15	11	4.68
	9	40	33	2.13
4	0	28	23	1.00
	1	7	5	32.14
	2	13	6	16.75
	3	10	7	12.90
	4	14	12	7.00
	5	13	9	6.18
	6	11	7	4.70
	7	17	12	3.31
	8	19	14	1.89
	9	22	16	1.82

Puis, pour déterminer l'effet de l'utilisation d'une covariable erronée dans le cas de la méthode cold deck simple, nous répétons les calculs qui précèdent en nous servant de la liste annuelle de paie de l'année courante comme covariable x et des revenus et de la liste de paie annuelle de l'année précédente pour les \bar{y} et \bar{x} , respectivement. Les résultats sont présentés au tableau 3.

Suit un résumé des résultats des tableaux 2 et 3.

1. Les résultats de la méthode cold deck simple dépendent fortement du choix de la covariable x . Si cette dernière correspond au revenu annuel de l'année précédente (tableau 2), l'écart entre les totaux estimatifs obtenus par les trois méthodes est négligeable; en ce qui concerne l'EQM estimative, la méthode cold deck simple est celle

Examinons maintenant l'estimation de la deuxième composante de la variance dans l'expression (13). Pour \hat{Y}_C ,

$$E_s(\hat{Y}_C^2) - Y^2 = \sum_{i \in \mathcal{P}} [a_i y_i + (1 - a_i) x_i] - \sum_{i \in \mathcal{P}} y_i - \sum_{i \in \mathcal{P}} (1 - a_i)(y_i - x_i).$$

Alors, en vertu du modèle (1),

$$V^m[E_s(\hat{Y}_C^2) - Y^2] =$$

$$E_m \left[\sum_{i \in \mathcal{P}} \sigma_k^2 (1 - a_i) x_i^2 + V^m \left[\sum_{i \in \mathcal{P}} (1 - a_i)(\beta_i - 1) x_i \right] \right].$$

Si nous estimons σ_k^2 au moyen de

$$\hat{\sigma}_2^k = \frac{\sum_{i \in \mathcal{S}_k} a_i w_i (y_i - \beta_k x_i)^2}{\sum_{i \in \mathcal{S}_k} a_i w_i x_i},$$

alors un estimateur de $V^m[E_s(\hat{Y}_C^2) - Y^2]$ prend la forme

$$v_{2C}^k = \sum_{i \in \mathcal{S}_k} \hat{\sigma}_2^k \sum_{i \in \mathcal{S}_k} (1 - a_i) w_i x_i +$$

$$\sum_{i \in \mathcal{S}_k} \frac{n_h}{N_h} \sum_{i \in \mathcal{S}_k(h)} \left(n_i' - \frac{1}{n_h} \sum_{i \in \mathcal{S}_k(h)} n_i' \right)^2,$$

(15)

$$\text{ou } n_i' = (1 - a_i)(\beta_i - 1) x_i \text{ et } \beta_i = \beta_k \text{ pour } i \in \mathcal{S}_k.$$

Pour \hat{Y}_{C-R} ,

$$E_s(\hat{Y}_{C-R}^2) - Y^2 = - \sum_{i \in \mathcal{P}} (1 - a_i)(y_i - x_i) y_i / x_i$$

et

$$V^m[E_s(\hat{Y}_{C-R}^2) - Y^2] = E_m \left[\sum_{i \in \mathcal{P}} \sigma_k^2 \sum_{i \in \mathcal{P}} (1 - a_i) x_i^2 + \sum_{i \in \mathcal{P}} \hat{\sigma}_2^k \sum_{i \in \mathcal{P}} (1 - a_i) x_i^2 / x_i \right].$$

Donc, on peut estimer $V^m[E_s(\hat{Y}_{C-R}^2) - Y^2]$ au moyen de

$$v_{2C-R}^k = \sum_{i \in \mathcal{S}_k} \left[\hat{\sigma}_2^k \sum_{i \in \mathcal{S}_k} (1 - a_i) w_i x_i + \hat{\sigma}_2^k \sum_{i \in \mathcal{S}_k} (1 - a_i) w_i x_i^2 / x_i \right], \quad (16)$$

où

$$\hat{\sigma}_2^k = \frac{\sum_{i \in \mathcal{S}_k} w_i (y_i - \beta_k x_i)^2}{\sum_{i \in \mathcal{S}_k} w_i x_i^2},$$

et

$$\hat{\beta}_k = \frac{\sum_{i \in \mathcal{S}_k} w_i y_i}{\sum_{i \in \mathcal{S}_k} w_i x_i}.$$

Pour \hat{Y}_R ,

$$E_s(\hat{Y}_R^2) - Y^2 \approx \sum_{i \in \mathcal{P}} \left[\sum_{i \in \mathcal{P}} x_i / \sum_{i \in \mathcal{P}} a_i x_i \right] \left[\sum_{i \in \mathcal{P}} a_i y_i - \sum_{i \in \mathcal{P}} y_i \right]$$

Donc, nous obtenons les erreurs quadratiques moyennes estimatives suivantes: $\text{eqm}(\hat{Y}_R)$ peut être estimée au moyen de

$$\widehat{\text{eqm}}(\hat{Y}_R) = v_{1R} + v_{2R},$$

où on calcule v_{1R} au moyen de (14) en posant $t_i = \zeta_i a_i (y_i - \beta_i x_i) + \beta_i x_i$, $\zeta_i = \zeta_k$ et $\beta_i = \beta_k$ pour $i \in \mathcal{S}_k$, et v_{2R} au moyen de l'expression (17); $\text{eqm}(\hat{Y}_C)$ est donné par

$$\widehat{\text{eqm}}(\hat{Y}_C) = v_{1C} + v_{2C} + \left[\sum_{i \in \mathcal{S}_k} (1 - \beta_k) \sum_{i \in \mathcal{S}_k} w_i x_i^2 \right],$$

où on obtient v_{1C} au moyen de (14) en posant $t_i = a_i y_i + (1 - a_i) x_i$, et v_{2C} au moyen de (15); $\text{eqm}(\hat{Y}_{C-R})$ peut être estimée par

$$\widehat{\text{eqm}}(\hat{Y}_{C-R}) = v_{1C-R} + v_{2C-R}.$$

où on obtient v_{1C-R} au moyen de (14) en posant $t_i = a_i y_i + (1 - a_i) x_i y_i / x_i$, et v_{2C-R} au moyen de (16).

Si l'on applique le modèle (1) et les paramètres asymptotiques de Krewski et Rao (1981), Rao et Shao (1992) ou Valliant (1993), les estimateurs de l'erreur quadratique moyenne que l'on dérive sont asymptotiquement sans biais et convergents à mesure que la taille de l'échantillon tend vers l'infini dans toutes les cellules d'imputation.

Dans le cas de l'imputation par la méthode cold deck ou de la méthode cold deck par quotient, le premier terme (v_{1C} ou v_{1C-R}) de l'erreur quadratique moyenne estimative est le même que celui obtenu en appliquant une formule type (comme (14)) et en traitant les valeurs imputées pour la non-réponse comme des données observées. Dans le cas de l'imputation par la méthode du quotient, si l'on applique la

et peut être estimé par

$$E_m \left\{ \sum_{i \in \mathcal{P}} \sigma_k^2 x_i \sum_{i \in \mathcal{P}} (1 - a_i) x_i \right\} / \sum_{i \in \mathcal{P}} a_i x_i^2$$

et, d'après le développement en série de Taylor,

$$V^m[E_s(\hat{Y}_R^2) - Y^2] \approx$$

$$v_{2R} = \sum_{i \in \mathcal{S}_k} \left[\sum_{i \in \mathcal{S}_k} w_i x_i^2 \sum_{i \in \mathcal{S}_k} (1 - a_i) w_i x_i \right] / \sum_{i \in \mathcal{S}_k} a_i w_i x_i. \quad (17)$$

Enfin, \hat{Y}_R et \hat{Y}_{C-R} sont des estimateurs non biaisés, mais \hat{Y}_C est entaché du biais

$$\sum_{i \in \mathcal{P}} (1 - \beta_k) E_m \left[\sum_{i \in \mathcal{P}} (1 - a_i) x_i \right],$$

que l'on peut estimer par

$$\sum_{i \in \mathcal{S}_k} (1 - \beta_k) \sum_{i \in \mathcal{S}_k} (1 - a_i) w_i x_i.$$

Donc, nous obtenons les erreurs quadratiques moyennes estimatives suivantes: $\text{eqm}(\hat{Y}_R)$ peut être estimée au moyen de

premier membre de (10), il suffit de remplacer $(p + 1/p)$ par y_i/μ_i .

Il faut estimer les paramètres β , σ , μ_x , v_x et y_x afin de comparer l'efficacité de \hat{Y}_R , \hat{Y}_C et \hat{Y}_{C-R} . Au lieu de cela, nous pouvons comparer directement les erreurs quadratiques moyennes estimatives de \hat{Y}_R , \hat{Y}_C et \hat{Y}_{C-R} . Nous examinons cette comparaison à la section qui suit.

3. ÉCHANTILLONNAGE STRATIFIÉ AVEC RÉPONSE SANS EFFET CONFOUSIONNEL

Considérons le plan de sondage stratifié qui suit adopté par nombre d'organismes gouvernementaux d'enquêtes aux Etats-Unis: la population finie P est stratifiée en H strates contenant N_h unités dans la h -ième strate; on sélectionne $n_h \geq 2$ unités sans remise dans la strate h , conformément à un plan d'échantillonnage probabiliste donné et les unités sont sélectionnées indépendamment dans les diverses strates.

Les poids de sondage w_j sont calculés de sorte que, si l'on observe tous les y_j , l'estimateur HT $\sum_{ies} w_j y_j$ est non biaisé pour Y dans les conditions d'échantillonnage répétée. Nous supposons que le modèle (1) est vérifié. La probabilité de réponse n'est plus constante, mais elle est indépendante de la valeur de y . Dans le cas de la méthode d'imputation cold deck par quotient, nous supposons aussi que, dans la k -ième cellule d'imputation, $y'_i = \beta_k x'_i + x'^{1/2}_i \bar{e}_i$, $E(\bar{e}_i) = 0$, $Var(\bar{e}_i) = \sigma_k^2$ et $e_i, \bar{e}_i, (x_i, x'_i)$ sont mutuellement

indépendants. L'estimation de l'erreur quadratique moyenne nous fondant sur les erreurs quadratiques moyennes estimatives, \hat{Y}_R , \hat{Y}_C et \hat{Y}_{C-R} est, en fait, un élément important de la théorie de l'échantillonnage. Il est bien connu que, pour les ensembles de données imputées, la méthode naïve consistant à appliquer les formules types d'estimation de la variance en traitant les valeurs imputées pour la non-réponse comme des données observées entraîne une sous-estimation. Si l'existence aucune méthode correcte (pour estimer l'erreur quadratique moyenne), nombre de bureaux d'enquête utilisent la méthode naïve.

Nous calculons maintenant les estimateurs de $Var(\hat{Y})$ ou $eqm(\hat{Y})$ qui sont corrects dans les conditions du modèle (1), où \hat{Y} représente \hat{Y}_R , \hat{Y}_C ou \hat{Y}_{C-R} .

Représentons par E_m et V_m l'espérance mathématique et la variance dans les conditions du modèle (1) et par E_s et V_s l'espérance mathématique et la variance dans les conditions d'échantillonnage répété (subordonnées au modèle et à la réponse). Alors

$$Var(\hat{Y} - Y) = E_m[Var_s(\hat{Y})] + V_m[E_s(\hat{Y}) - Y]. \quad (13)$$

Shao: Imputation par la méthode cold deck et la méthode du quotient

$$v_1 = \sum_{i=1}^h \left(1 - \frac{N_h}{n_h} \right) \frac{n_h - 1}{n_h} \sum_{ies(h)} \left(w_{i,t'_i} - \frac{1}{n_h} \sum_{ies(h)} w_{i,t'_i} \right)^2 \quad (14)$$

Supposons que

avantages de l'utilisation de la méthode cold deck.

L'estimation de $Var_s(\hat{Y}_C)$ et $Var_s(\hat{Y}_{C-R})$ est simple (l'un des avantages de l'utilisation de la méthode cold deck).

pour Y'_i . L'estimation de $Var_s(\hat{Y}_R)$ est simple (l'un des avantages de l'utilisation de la méthode cold deck du quotient), où a'_i est l'indicateur de réponse pour Y'_i . L'estimation de $Var_s(\hat{Y}_{C-R})$ est simple (l'un des avantages de l'utilisation de la méthode cold deck du quotient), où a'_i est l'indicateur de réponse pour Y'_i . L'estimation de $Var_s(\hat{Y}_{C-R})$ est simple (l'un des avantages de l'utilisation de la méthode cold deck du quotient), où a'_i est l'indicateur de réponse pour Y'_i .

est l'estimateur type de la variance de $\sum_{ies} w_{i,t'_i}$ si l'on traite $\{t'_i, i \in s\}$ comme un échantillon observé (tiré de $\{t'_i, i \in P\}$), où $s(h)$ est la valeur de s limitée à la strate h . Alors, on peut estimer $Var_s(\hat{Y}_{C-R})$ au moyen de l'expression (14) en posant que $t'_i = a'_i y'_i + (1 - a'_i) x'_i$ et $t'_i = a'_i y'_i + (1 - a'_i) x'_i$. L'estimation de $Var_s(\hat{Y}_R)$ est un peu plus compliquée, mais semblable. Supposons que, dans chaque cellule d'imputation, le nombre d'unités échantillonnées est grand et que les probabilités de réponse sont bornées de façon à ce qu'elles ne soient pas nulles. Remarquons que

$$\hat{Y}_R = \sum_{ies}^k \left[\left(\sum_{ies}^k w_{i,x'_i} / \sum_{ies}^k w_{i,x'_i} \right) \left(\sum_{ies}^k w_{i,y'_i} - \beta_k x'_i \right) + \beta_k \sum_{ies}^k w_{i,x'_i} \right] \times \left[\sum_{ies}^k w_{i,y'_i} - \beta_k x'_i \right] + \beta_k \sum_{ies}^k w_{i,x'_i}$$

où $\zeta_k = E(\sum_{ies}^k w_{i,x'_i}) / E(\sum_{ies}^k w_{i,y'_i})$ et $\zeta'_i = \zeta_k$ et $\beta'_i = \beta_k$ pour $i \in s_k$. Après avoir estimé β_k au moyen de β'_k et ζ_k au moyen de $\zeta'_k = \sum_{ies}^k w_{i,x'_i} / \sum_{ies}^k w_{i,y'_i}$, nous estimons $Var_s(\hat{Y}_R)$ au moyen de l'expression (14) en posant $t'_i = \zeta'_i a'_i (y'_i - \beta'_i x'_i) + \beta'_i x'_i$ et $\zeta'_i = \zeta_k$ et $\beta'_i = \beta_k$ pour $i \in s_k$. Avant d'examiner l'estimation de $Var_s(\hat{Y}_R)$, qui est le deuxième élément de la variance dans l'expression (13), soulignons que $Var_s(\hat{Y}_R) - Y$ / $E_m[Var_s(\hat{Y}_R)] = O(n/N)$, parce que la variance de $E_s(\hat{Y}_R) - Y$ (si elle n'est pas nulle) est ordinairement d'ordre N^2/n et, donc, que $Var_s(\hat{Y}_R)$ est habituellement d'ordre N^2/n dans certaines conditions de régularité. Par conséquent, en théorie, il n'est pas nécessaire d'estimer $Var_s(\hat{Y}_R)$ [si la fraction d'échantillonnage n/N est négligeable. Cependant, comme on ne connaît pas le terme constant de $O(n/N)$, on pourrait encore vouloir estimer $Var_s(\hat{Y}_R) - Y$ dans les applications, même si n/N est faible

cellule d'imputation, si bien que nous pouvons abandonner l'indice k pour la cellule d'imputation; enfin, la probabilité de réponse est une constante $p > 0$ (mécanisme de réponse uniforme).

Dans ces conditions, $w_i = N/n$, où n est l'effectif de l'échantillon et N est l'effectif de la population \mathcal{P} . Puisque l'on suppose que $n/N \approx 0$,

$$\text{eqm}(\hat{Y}^R) \approx N^2 \left(\frac{u}{\sigma^2 \mu_x} + \beta^2 v_x \right) \quad (7)$$

pour une valeur élevée de n , où $\mu_x = E(x_i)$ et $v_x = V(x_i)$ et, dans tout l'article, $A \approx B$ signifie que A est égal à B jusqu'à un terme qui est relativement négligeable comparativement à A et B , puisque dans toutes les cellules d'imputation, l'effectif de l'échantillon tend vers l'infini. Une explication plus détaillée de la dérivation du résultat (7) figure en annexe. D'après \hat{Y}^w dans l'équation (4), il est facile de voir que $w_i = N/r$, où r est la taille de r et \hat{Y}^w n'est pas biaisé.

Par conséquent, \hat{Y}^R est plus efficace que \hat{Y}^w , à moins que $p = 1$ et $\beta^2 v_x = 0$. Le gain réalisé en se servant de \hat{Y}^R est proportionnel à β^2 et à v_x , qui sont deux mesures de l'utilité de la variable auxiliaire x pour expliquer y conformément au modèle (1).

$$\hat{Y}^C = \frac{n}{N} \left(\sum_{i \in r} y_i + \sum_{i \in s} x_i \right) = \frac{n}{N} \left(\sum_{i \in r} x_i e_i + \beta \sum_{i \in s} x_i + \sum_{i \in s} x_i \right),$$

où les e_i sont définis en (1). Par conséquent,

$$V(\hat{Y}^C) = \frac{n}{N^2} \left\{ \sigma^2 d \mu_x + (\beta^2 d^2 + 1 - d) v_x \right. \\ \left. + (\beta - 1)^2 d^2 (1 - d) \mu_x^2 \right\} \quad (8)$$

(voir l'annexe). Le biais de \hat{Y}^C est

$$E(\hat{Y}^C) - Y = N \mu_x (1 - d) (1 - \beta)$$

et, donc,

$$\text{eqm}(\hat{Y}^C) = V(\hat{Y}^C) + [E(\hat{Y}^C) - Y]^2 \\ \approx V(\hat{Y}^C) + [E(\hat{Y}^C) - Y]^2$$

$$= N^2 \left\{ \sigma^2 d \mu_x + (\beta^2 d^2 + 1 - d) v_x \right.$$

$$\left. + (\beta - 1)^2 (1 - d) [d + u(1 - d)] \mu_x^2 \right\}. \quad (9)$$

La comparaison de (7) et (9) mène aux conclusions suivantes:

1. si $p = 1$ (non-réponse), $\text{eqm}(\hat{Y}^C) = \text{eqm}(\hat{Y}^R)$;
2. si $p < 1$ et $\beta = 1$ (v et x ont la même moyenne), $\text{eqm}(\hat{Y}^C) < \text{eqm}(\hat{Y}^R)$;
3. si $p < 1$ et $\beta \neq 1$, $\text{eqm}(\hat{Y}^C) \leq \text{eqm}(\hat{Y}^R)$ si, et uniquement si

$$(\beta - 1)^2 [d^2 p + n(1 - d)] \mu_x + (1 - \beta^2) v_x / \mu_x$$

$$- \sigma^2 (d + 1) / d \leq 0. \quad (10)$$

Supposons que $\mu_x > 0$. Dans le cas de la plupart des enquêtes économiques, la variance relative v_x / μ_x^2 est plus faible que $d + n(1 - d)$. Par conséquent, le premier membre de (10) est une fonction quadratique de β dont le coefficient du terme β^2 est positif et, par conséquent, la méthode cold deck simple donne de meilleurs résultats si β est compris dans l'intervalle dont les limites sont

$$\frac{d + n(1 - d) [\mu_x^2 v_x^2 / \mu_x^2 + \{ [d + n(1 - d)] \mu_x - v_x / \mu_x \} \sigma^2 (d + 1) / d]}{[d + n(1 - d)] \mu_x - v_x / \mu_x}.$$

Cet intervalle contient la valeur 1 puisque (10) est vérifié si $\beta = 1$. À noter que $[d + n(1 - d)] \mu_x$ tend vers l'infini quand n tend vers l'infini. Par conséquent, l'intervalle des β pour lesquels la méthode cold deck simple donne de meilleurs résultats se réduit à un point unique si $(\beta = 1)$ quand $n \rightarrow \infty$.

Examinons maintenant la méthode cold deck par quotient. Supposons que $y_i = \beta x_i + x_i^{1/2} e_i$, $E(e_i) = 0$, $V(e_i) = \sigma^2$, et que e_i, x_i et $(x_i, x_i^{1/2})$ sont mutuellement indépendants. Soit $z_i = x_i y_i / x_i^{1/2}$ et $e'_i = y_i - z_i = x_i^{1/2} e_i$.

$$\text{eqm}(\hat{Y}^{C-R}) = \frac{n}{N^2} \left\{ \sigma^2 d \mu_x + \beta^2 v_x + \sigma^2 (1 - d) v_x \right. \\ \left. \right\} \quad (11)$$

où $y_x = E(x_i^{1/2} / x_i)$ (voir l'annexe). D'après (7) et (11),

$$\text{eqm}(\hat{Y}^R) - \text{eqm}(\hat{Y}^{C-R}) = \frac{n}{N^2 \sigma^2 (1 - d)} \left\{ \left(\frac{d}{1} + 1 \right) \mu_x - y_x \right\} \quad (12)$$

et, par conséquent, la méthode cold deck par quotient donne de meilleurs résultats que la méthode d'imputation par quotient si, et uniquement si $1/d + 1 \geq y_x / \mu_x$. À noter que $y_x \geq \mu_x$ et x_i s'approche de μ_x et $x_i^{1/2}$ sont fortement positivement corrélés, auquel cas la méthode d'imputation cold deck par quotient peut être nettement supérieure à la méthode cold deck simple et la comparaison entre la méthode cold deck simple et la méthode cold deck par quotient est identique à la comparaison entre la méthode cold deck simple et la méthode d'imputation par quotient. Dans le troisième terme du

$$\hat{Y}^w = \sum_{i \in R_k} w_i^k Y_i^w, \quad w_i^k = w_i / \left(\sum_{i \in R_k} w_i \right) \quad (4)$$

On constate que, si $x_i = 1$, alors les estimateurs représentés par les équations (3) et (4) sont les mêmes. Les deux estimateurs sont dépourvus de biais si le modèle (1) est vérifié. (Dans le présent article, le biais et la variance ont trait au modèle (1) et à l'échantillonnage répété, sauf indication contraire.) Cependant, dans les conditions du modèle (1), \hat{Y}^R est plus efficace que \hat{Y}^w si r est de taille nettement plus faible que s . Même si le modèle de régression décrit en (1) n'est pas vérifié, \hat{Y}^R peut encore être plus efficace que \hat{Y}^w si l'on s'en tient à l'erreur quadratique moyenne en cas d'échantillonnage répété (Cochran 1977, chapitre 6) quand la probabilité de réponse est une constante dans toute cellule d'imputation (condition qui assure que \hat{Y}^R et \hat{Y}^w soient approximativement non biaisés en ce qui concerne l'échantillonnage répété).

L'objet de la présente note est de comparer l'efficacité de \hat{Y}^R à celle d'autres estimateurs de Y d'après les données obtenues par imputation de valeurs aux cas de non-réponse par une méthode appelée cold deck. La méthode d'imputation cold deck consiste à imputer aux cas de non-réponse pour la variable Y des valeurs déclarées pour toute autre variable que Y (par exemple, valeurs déclarées pour une covariable et/ou) provenant d'une autre enquête). L'imputation par la méthode cold deck est l'opposé de l'imputation par la méthode hot deck en vertu de laquelle on impute à un cas de non-réponse une valeur déclarée pour la même variable durant l'enquête courante. La méthode d'imputation par quotient consiste à utiliser à la fois des valeurs déclarées de Y et des données auxiliaires et est parfois appelée méthode «warm deck». La forme la plus simple de méthode d'imputation cold deck consiste à imputer à une non-réponse y_i , $i \in s - r$, une valeur x_i , l'estimateur HT résultant de Y s'écrivant alors:

$$\hat{Y}^C = \sum_{i \in s} w_i' y_i' + \sum_{i \in s-r} w_i' x_i' \quad (5)$$

L'utilisation de cette méthode cold deck simple est motivée par le fait que, en vertu du modèle (1), la valeur des β_k est proche de 1 dans nombre de problèmes d'enquête, particulièrement quand les x_i sont des valeurs de Y provenant d'une enquête précédente. Si certains β_k ne sont pas égaux à 1, dans l'équation (5), \hat{Y}^C est entraîné d'un biais qui ne disparaît pas, même si $s = \mathcal{P}$ (c'est-à-dire si l'échantillon correspond à un dénombrement complet (recensement)). Cependant, le fait qu'un léger biais persiste peut être compensé par la diminution de la variance si bien que l'erreur quadratique moyenne $\text{eqm}(\hat{Y}^C) = E(\hat{Y}^C - Y)^2$ demeure plus faible que l'erreur quadratique moyenne $\text{eqm}(\hat{Y}^R) = E(\hat{Y}^R - Y)^2 = E(\hat{Y}^R - Y)^2$, où E et V représentent l'espérance mathématique et la variance dans les conditions du modèle (1) et de l'échantillonnage répété. Des précisions supplémentaires sont données à la section 2. On peut

améliorer l'imputation en passant de la méthode cold deck simple à la méthode cold deck par quotient, qui consiste à imputer pour une non-réponse y_i une valeur $x_i y_i' / x_i'$, où y_i' et x_i' sont des valeurs qui ont été déclarées lors d'une enquête précédente. L'estimateur HT correspondant de Y prend la forme

$$\hat{Y}^{C-R} = \sum_{i \in r} w_i' y_i' + \sum_{i \in s-r} w_i' x_i y_i' / x_i' \quad (6)$$

L'estimateur décrit en (6) est sans biais si le modèle (1) est vérifié pour y_i' et x_i' (c'est-à-dire $y_i' = \beta_k x_i' + x_i'^{1/2} e_i'$) en se servant du même β_k que pour y_i et x_i . Le U.S. Census Bureau (King et Kombaru 1994) et le U.S. Bureau of Labor Statistics (Butani, Harter et Wolter 1998) appliquent ces deux méthodes cold deck à grande échelle dans le cas des enquêtes de nature économique. L'application des méthodes cold deck d'imputation n'exige pas que l'on connaisse les cellules d'imputation, mais le modèle (1) suppose qu'aucun biais n'entraîne \hat{Y}^C ni \hat{Y}^{C-R} .

Bien qu'elle s'appuie sur un plus grand nombre de données auxiliaires, la méthode cold deck par quotient ne donne pas nécessairement de meilleurs résultats que la méthode cold deck simple ou que la méthode d'imputation par quotient. À la section 2, on compare explicitement les erreurs quadratiques moyennes de \hat{Y}^R , \hat{Y}^C et \hat{Y}^{C-R} dans le cas particulier où l'échantillon s est un échantillon aléatoire simple (EAS) et où la probabilité de réponse est constante. On détermine les conditions dans lesquelles une méthode est supérieure aux autres. Si le plan d'échantillonnage ou le mécanisme de réponse est complexe, il est difficile de comparer explicitement les erreurs quadratiques moyennes. Cependant, on peut estimer l'erreur quadratique moyenne de \hat{Y}^R , \hat{Y}^C et \hat{Y}^{C-R} et faire une comparaison empirique. L'estimation de la variance ou de l'erreur quadratique moyenne est, en soi, un problème important, puisqu'il est courant de déclarer les estimations de la variance ou de l'erreur quadratique moyenne en même temps que les totaux estimatifs. Cette question est abordée à la section 3. Nos résultats peuvent aussi être appliqués dans les conditions de l'échantillonnage à deux degrés ou de l'échantillonnage double, que l'on emploie souvent quand la sélection d'un grand échantillon $\{x_i', i \in s\}$ est peu coûteuse mais qu'il est onéreux d'obtenir des valeurs de y_i si bien qu'on sélectionne un sous-échantillon $\{y_i', i \in r\}$ à la deuxième phase, $r \subset s$.

Un exemple numérique axé sur les données de l'enquête annuelle sur les transports réalisés par le U.S. Census Bureau est examiné à la section 4.

2. EAS AVEC RÉPONSE UNIFORME

Pour illustrer l'idée, commençons par exposer le cas le plus simple où s est un échantillon aléatoire simple (EAS) (sans remise provenant de \mathcal{P} , mais dont la fraction d'échantillonnage est négligeable); il n'existe qu'une seule

Imputation par la méthode du quotient

JUN SHAO¹

RÉSUMÉ

L'imputation est une méthode utilisée couramment pour compenser l'effet de la non-réponse lors de l'analyse des données d'enquête. Fondée sur des données auxiliaires, l'imputation peut produire des estimateurs plus efficaces que ceux construits en ne tenant compte ni de la non-réponse ni de la reproduction. Nous étudions et comparons l'erreur quadratique moyenne d'estimateurs d'enquête fondés sur des données imputées par trois méthodes distinctes, c'est-à-dire la méthode courante d'imputation par quotient et deux méthodes cold deck fréquemment appliquées aux enquêtes de nature économique réalisées par le U.S. Census Bureau et par le U.S. Bureau of Labor Statistics. La méthode cold deck consiste à imputer à une non-réponse à une question particulière toute autre valeur que celles déclarées pour la même question de l'ensemble de données courantes (par exemple, valeur observée pour une covariable et/ou) tirée d'une enquête antérieure). Bien qu'elle s'appuie sur un plus grand nombre de données auxiliaires que les autres méthodes, l'imputation par la méthode cold deck ne donne pas nécessairement de meilleurs résultats si l'on s'en tient à l'erreur quadratique moyenne des estimateurs d'enquête résultants. À l'aide d'un exemple simple, nous comparons explicitement les erreurs quadratiques moyennes et déterminons dans quelles conditions l'une des méthodes est supérieure aux autres. Pour les cas généraux, nous proposons de comparer empiriquement les erreurs quadratiques moyennes en se fondant sur certaines estimations cohérentes de l'erreur quadratique moyenne. L'estimation de l'erreur quadratique moyenne des estimateurs d'enquête en cas d'imputation de données est en soi un problème important. À titre d'illustration, nous présentons un exemple numérique basé sur l'enquête annuelle sur les transports.

MOTS CLÉS: Enquête complexe; erreur quadratique moyenne; non-réponse; échantillonnage aléatoire simple; estimation de la variance.

1. INTRODUCTION

L'imputation est l'une des méthodes les plus courantes utilisées pour compenser la non-réponse lors de l'analyse des données d'enquête. En plus des nombreuses raisons pratiques qui la justifient, l'imputation au moyen de données auxiliaires peut produire des estimateurs plus efficaces que ceux construits en ne tenant compte ni de la non-réponse ni de la reproduction. Supposons que l'on tire un échantillon s d'une population finie P et qu'il comprenne certaines unités représentées par $i = 1, \dots, M$, et que l'on observe $\{y_i, i \in r\}$ (répondants), $r \subset s$. Supposons en outre que l'on dispose de données auxiliaires x_i observées pour tous les $i \in s$ et $x_i > 0$. La méthode d'imputation par quotient utilise couramment (voir, par exemple, Kalton et Kasprzyk 1986) consiste à imputer des valeurs pour la non-réponse comme suit. En premier lieu, on crée K cellules d'imputation $P_1^k \cup P_2^k \cup \dots \cup P_K^k = P$, d'après une variable auxiliaire nominale (que l'on observe pour chaque $i \in s$ et qui diffère typiquement de x) telles que, pour chaque k , on suppose qu'est vérifié le modèle suivant:

$$y_i = \beta_k x_i + x_i^{1/2} e_i, \quad i \in P_k^k \quad (1)$$

$$P(a_i = 1 | y_i, x_i) = P(a_i = 1 | x_i),$$

où β_k est un paramètre inconnu, e_i est indépendante de x_i , avec $E(e_i) = 0$ et $V(e_i) = \sigma_k^2 > 0$ inconnu, a_i indique si y_i est un répondant et (a_i, x_i) sont indépendants. Alors, dans

la cellule d'imputation k , on impute au non-répondant y_i la valeur $\beta_k x_i$, où

$$\beta_k = \sum_{i \in r_k} w_i y_i / \sum_{i \in r_k} w_i x_i \quad (2)$$

est le meilleur estimateur linéaire non biaisé de β_k étant donné le modèle (1), r_k est r limité à la k -ième cellule d'imputation et w_i est le poids de sondage associé à la i -ième unité échantillonnée. À noter que le modèle (1) est un modèle de régression de y_i sur x_i (sans coordonnée à l'origine et avec variance de l'erreur proportionnelle à x_i) et un modèle de réponse fondé sur l'hypothèse selon laquelle le mécanisme de réponse est indépendant des y_i , étant donné les x_i . Ce mécanisme de réponse est qualifié de réponse à données aléatoirement manquantes par Rubin (1976) ou de réponse sans effet confusional par Lee, Rancourt et Särndal (1994). D'après l'ensemble de données imputées, l'estimateur d'Horvitz-Thompson (HT) de X , c'est-à-dire le total de population des y_i , s'écrit

$$\hat{X}_r = \sum_{i \in r_k} \left(\sum_{j \in r_k} w_j y_j + \sum_{j \in s_k - r_k} w_j \beta_k x_j \right) \quad (3)$$

où s_k est la valeur de s limité à la k -ième cellule d'imputation. L'estimateur HT de X , calculé sans tenir compte de la non-réponse ni de la reproduction dans chaque cellule d'imputation, est représenté par

$$\left(\frac{\sigma^2}{\sigma^2 + \sigma_v^2} \right)^4 \{ (\mu_i - x_i' \beta)^2 - \sigma_v^2 \}^2,$$

tandis que l'erreur quadratique moyenne $\text{mse}_c(\hat{\mu}_i)$ ne comporte qu'une composante de la variance donnée par

$$\left(\frac{\sigma^2}{\sigma^2 + \sigma_v^2} \right)^4 \text{Var}_S \{ (y_i - x_i' \beta)^2 \}$$

$$= \left(\frac{\sigma^2}{\sigma^2 + \sigma_v^2} \right)^4 \{ 2\sigma_u^2 + 4(\mu_i - x_i' \beta)^2 \sigma_v^2 \}.$$

On peut évaluer l'efficacité signalée dans la proposition 4 comme le ratio des deux fluctuations moyennes définies ci-dessus. Elle est donnée par

$$\frac{2\sigma^4 + 4\sigma^2 \sum (\mu_i - x_i' \beta)^2 / n}{\sum \{ (\mu_i - x_i' \beta)^2 - \sigma_v^2 \}^2 / n}.$$

Le fait de prendre les espérances du numérateur et du dénominateur relativement au modèle (6) fournit le résultat.

BIBLIOGRAPHIE

- BARNDORFF-NIELSEN, O.E. et COX, D.R. (1989). *Asymptotic Techniques for Use in Statistics*. New York: Chapman and Hall.
- BELMONTÉ, E. (1998). Estimation dans les petites régions: une nouvelle définition de l'erreur quadratique moyenne de Prasad-Rao. *Recueil 1998 de la Section des méthodes d'enquête, Société Statistique du Canada*, 165-170.
- BELMONTÉ, E. (1999). *L'estimation dans les petites régions: Survol des méthodes de Bayes et présentation d'un estimateur conditionnel de l'EQM*. Mémoire de maîtrise. Département de mathématiques et de statistique, Université Laval.
- BILODEAU, M., et SRIVASTAVA, M.S. (1988). Estimation of the MSE matrix of the Stein estimator. *La revue canadienne de Statistique*, 16, 153-159.
- BOOTH, J.G., et HOBERT, J.P. (1998). Standard errors of predictions in generalized linear mixed models. *Journal of the American Statistical Association*, 93, 262-272.
- BURGESS, R.D. (1988). Évaluation des estimations du sous-recensement du Canada. *Techniques d'enquête*, 14, 147-167.
- CRESSIE, N. (1992). MVC dans le lissage du taux sous-dénombrement du recensement selon l'approche empirique de Bayes. *Techniques d'enquête*, 18, 83-103.
- DICK, P. (1995). Modélisation du sous-dénombrement dans le recensement du Canada 1991. *Techniques D'enquête*, 21, 51-61.
- FAY, R.E., et HERRIOT, R.A. (1979). Estimates of income for small places: An application of James Stein procedure to census data. *Journal of the American Statistical Association*, 74, 269-277.
- GHOSH, M., et RAO, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-93.
- HOGAN, H. (1992). The 1990 Post-Enumeration Survey: an overview. *The American Statistician*, 46, 261-269.
- KACKAR, R.N., et HARVILLE, D.A. (1984). Approximations for standard errors of estimators for fixed and random effects in mixed models. *Journal of the American Statistical Association*, 79, 853-862.
- KOTT, P.S. (1989). Estimation robuste pour petits domaines à l'aide du modèle des effets aléatoires. *Techniques d'enquête*, 15, 3-13.
- LAHIRI, P., et RAO, J.N.K. (1995). Robust estimation of mean squared error of small area estimators. *Journal of the American Statistical Association*, 90, 758-766.
- MARTZ, J.S., et LWIN, T. (1989). *Empirical Bayes Methods*. (Deuxième édition), London: Chapman and Hall.
- PRASAD, N.G.N., et RAO, J.N.K. (1990). The estimation of mean squared errors of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.
- PRASAD, N.G.N., et RAO, J.N.K. (1999). Estimation régionale robuste au moyen d'un modèle simple à effets aléatoires. *Techniques d'enquête*, 25, 73-79.
- PURCELL, N.J., et KISH, L. (1979). Estimation for small domains. *Biometrics*, 35, 365-384.
- RAO, C.R., et SHINOZAKI, N. (1978). Precision of individual estimators in simultaneous estimation of parameters. *Biometrika*, 65, 23-30.
- REID, N. (1991). Approximations and asymptotics. Dans *Statistical Theory and Modeling. In Honor of Sir David Cox*, FRS, (éds. D.V. Hinkley, N. Reid et E.J. Snell), 287-305.
- RIVEST, L.P. (1995). A composite estimator for provincial under-coverage in the Canadian census. *Recueil 1995 de la Section des méthodes d'enquête, Société Statistique du Canada*, 33-38.
- ROBERT, C. (1992). *L'Analyse Statistique Bayésienne*. Paris: Economica.
- ROYCE, D. (1992). Une comparaison d'estimateurs d'un ensemble de totaux de population. *Techniques d'enquête*, 18, 121-138.
- SÄRNDALE, C.E., SWENSSON, B., et WREFTMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- SCOTT, A., et SMITH T.M.F. (1971). Interval estimates for linear combinations of means. *Applied Statistics*, 20, 276-285.
- SINGH, M.P., GAMBINO, J., et MANTTEL, H.J. (1994). Les petites régions: problèmes et solutions. *Techniques d'enquête*, 20, 3-23.
- SINGH, A.C., STUKEL, D.M., et PFEFFERMANN, D. (1998). Bayesian versus frequentist measures of error in small area estimation. *Journal of the Royal Statistical Society B*, 60, 377-396.
- STEIN, C. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9, 1135-1151.
- WOLTER, K.M. (1985). Introduction to Variance Estimation. New York: Springer-Verlag.

Puisque $\Sigma_{1/2}$ est symétrique, $\Sigma_{1/2}^H = \Sigma_{1/2}$. Ainsi l'expression ci-dessus est le produit scalaire de e_j , la i -ième ligne de $\Sigma_{1/2}$, représente un vecteur $n \times 1$ de 0 sauf pour la i -ième composante qui est 1, et $\Delta g_j(\mu + \Sigma_{1/2} z) e_j$, la j -ième colonne de $\Delta g_j(\mu)$, évaluée en $y = \mu + \Sigma_{1/2} z$. Nous avons

$$E\{z_i g_j(\mu + \Sigma_{1/2} z) \mid z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n\} = e_j' \Sigma_{1/2} E\{\Delta g_j(\mu + \Sigma_{1/2} z) e_j \mid z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n\}$$

Cette égalité s'applique également de façon inconditionnelle, $E\{z_i g_j(\mu + \Sigma_{1/2} z)\} = e_j' \Sigma_{1/2} E\{\Delta g_j(\mu + \Sigma_{1/2} z)\} e_j$.

Autrement dit,

$$E\{zg(\mu + \Sigma_{1/2} z)\} = \Sigma_{1/2} E\{\Delta g(\mu + \Sigma_{1/2} z)\}.$$

CQFD.

Preuve de la proposition 2

Soit E_j l'espérance conditionnelle relativement à y_j , étant donné $(y_1, \dots, y_{j-1}, y_{j+1}, \dots, y_n)$, pour des valeurs fixes de $(y_1, \dots, y_{j-1}, y_{j+1}, \dots, y_n)$. On a

$$E_j\{y_j - \mu_j\} h(y_j) = \int^R (t - \mu_j) h(t) f(t) dt.$$

Pour évaluer cette expression, il est possible d'intégrer par parties. L'intégration de $(t - \mu_j) \exp\{-(t - \mu_j)^2 / (2\sigma_j^2)\}$ dans la fonction à intégrer donne

$$E_j\{y_j - \mu_j\} h(y_j) = \sigma_j^2 E_j\{h'(y_j)\} + \frac{2}{\sigma_j^{1/2}} \int^R h(t) \left\{ \frac{\sigma_j^2}{(t - \mu_j)^2} - 1 \right\} \frac{(2\pi\sigma_j^2)^{1/2}}{\exp\{(t - \mu_j)^2 / (2\sigma_j^2)\}} dt,$$

où $h'(t)$ est la dérivée de $h(t)$. Des intégrations répétées par parties indiquent que

$$\int^R h(t) \frac{\sigma_j^2}{(t - \mu_j)^2} \exp\{(t - \mu_j)^2 / (2\sigma_j^2)\} (2\pi\sigma_j^2)^{1/2} dt = \int^R \{h'(t)(t - \mu_j) + h(t)\} \frac{(2\pi\sigma_j^2)^{1/2}}{\exp\{(t - \mu_j)^2 / (2\sigma_j^2)\}} dt = \int^R \{\sigma_j^2 h''(t) + h(t)\} \frac{(2\pi\sigma_j^2)^{1/2}}{\exp\{(t - \mu_j)^2 / (2\sigma_j^2)\}} dt$$

où $h''(t)$ est la dérivée seconde de $h(t)$. Cela donne

$$E_j\{y_j - \mu_j\} h(y_j) = \sigma_j^2 E_j\{h''(y_j)\} + \frac{2}{\sigma_j^{3/2}} E_j\{h'(y_j)\} + o(p_j).$$

Soit E_j l'espérance prise relativement à la distribution de \hat{y}_j , étant donné toutes les autres quantités aléatoires $(y, \hat{y}_{j/f} \neq t)$. Dans ce contexte on peut écrire $(\partial g_j(y)) / (\partial y_j) = h(\hat{y}_j)$, où h est une fonction qui dépend possiblement de $(y, \hat{y}_{j/f}, j \neq t)$. Un développement en série de Taylor pour $h(y, \hat{y}_{j/f}, j \neq t)$ donne:

$$h(\hat{y}_j) = h(\sigma_j^2) + h'(\sigma_j^2)(\hat{y}_j - \sigma_j^2) + h''(\sigma_j^2) \frac{(\hat{y}_j - \sigma_j^2)^2}{2} + O((\hat{y}_j - \sigma_j^2)^3).$$

Puisque $(k - 1)\hat{\sigma}_j^2 / \sigma_j^2$ comporte une distribution χ_{k-1}^2 , $E_j\{(\hat{y}_j - \sigma_j^2)^2\} = 2\sigma_j^2 / (k - 1)$, et les moments centrés d'ordre supérieur sont $O(1/k^2)$. Le développement ci-dessus se ramène à

$$\sigma_j^2 E_j\{\partial g_j(y) / \partial y_j\} = \sigma_j^2 h(\sigma_j^2) + h''(\sigma_j^2) \frac{k-1}{3} + O(1/k^2)$$

Il est clair que le biais de $\hat{\sigma}_j^2 h(\hat{\sigma}_j^2)$ à titre d'estimateur de cette expression est de $O(1/k)$, pourvu que $h'(\sigma_j^2) \neq 0$. En laissant de côté les termes $O(1/k^2)$, on a

$$\left\{ \hat{\sigma}_j^2 h'' \left(\frac{k+1}{(k-1)\hat{\sigma}_j^2} \right) E_j \right\} \approx \sigma_j^2 h(\sigma_j^2) + h'(\sigma_j^2) E_j \left\{ \hat{\sigma}_j^2 \left(\frac{k+1}{(k-1)\hat{\sigma}_j^2} - \sigma_j^2 \right) \right\}$$

$$+ \frac{2}{h''(\sigma_j^2)} E_j \left\{ \hat{\sigma}_j^2 \left(\frac{k+1}{(k-1)\hat{\sigma}_j^2} - \sigma_j^2 \right) \right\}^2$$

Des opérations élémentaires indiquent que, dans la formule ci-dessus, le coefficient de $h'(\sigma_j^2)$ est nul et que

$$E_j \left\{ \hat{\sigma}_j^2 \left(\frac{k+1}{(k-1)\hat{\sigma}_j^2} - \sigma_j^2 \right) \right\}^2 = 2 \frac{k-1}{3} \sigma_j^2 + O(1/k^2).$$

Cela montre que

$$E_j \left\{ \hat{\sigma}_j^2 h'' \left(\frac{k+1}{(k-1)\hat{\sigma}_j^2} \right) \right\} = \sigma_j^2 E_j \{\partial g_j(y) / \partial y_j\} + O(1/k^2).$$

On complète la preuve en notant que cette égalité s'applique à l'espérance inconditionnelle, prise relativement à la distribution mixte de $(y, \hat{y}_j, j = 1, \dots, n)$.

Preuve de la proposition 4

L'erreur quadratique moyenne de la variance à posteriori à titre d'estimateur de l'erreur quadratique moyenne conditionnelle comporte uniquement un terme de biais,

6. CONCLUSIONS

L'estimateur de l'erreur quadratique moyenne conditionnelle proposée dans le présent exposé comporte plusieurs aspects intéressants. Il peut être mis en œuvre à l'aide de toute stratégie de réduction. Il est conditionnel en ce sens qu'il dépend de la réalisation du modèle de lissage servant à produire les caractéristiques régionales; l'estimateur conditionnel comporte donc une grande variance d'échantillonnage. De simples modifications de l'estimateur permettent de traiter l'asymétrie des données et des variances estimatives. Dans un contexte empirique de Bayes, il fournit des renseignements diagnostiques au sujet du modèle de lissage. Il peut également servir d'élément pour la construction d'estimateurs de la variance liée au modèle prédictif lorsque celle-ci ne trouve aucune expression de forme analytique.

REMERCIEMENTS

Nous tenons à remercier Peter Dick d'avoir fourni l'ensemble de données analysé à la section 5.2, et Jon Rao d'avoir signalé l'instabilité de l'estimateur conditionnel en présence de lissage poussé. L'appui financier du Fonds pour la formation des chercheurs et l'aide à la recherche du Québec et du Conseil de recherches en sciences naturelles et en génie du Canada est vivement apprécié.

ANNEXE

Preuve de la proposition 1

Soit $\Sigma_{1/2}$ une racine carrée symétrique pour Σ , telle que une distribution $N''(0, I)$. Pour ce qui est du vecteur aléatoire $z, E\{(\nu - \mu)g(\nu')\} = \Sigma_{1/2}E\{zg(\mu + \Sigma_{1/2}z)\}$. Ainsi l'espérance conditionnelle de $g_j(\mu + \Sigma_{1/2}z)$ donnée par $(z_1', \dots, z_{l-1}', z_{l+1}', \dots, z_n')$ est égale à

$$\int_R z' \exp(-z'^2/2) \frac{g_j(\mu + \Sigma_{1/2}z) dz_{l'}}{g_j(\mu + \Sigma_{1/2}z)}$$

Une intégration par parties indique que l'intégrale ci-dessus est égale à

$$\int_R \frac{\sqrt{2\pi}}{\exp(-z'^2/2)} \frac{\partial g_j(\mu + \Sigma_{1/2}z)}{\partial z_{l'}} dz_{l'}$$

Nous observons que

$$\frac{\partial z_{l'}}{\partial g_j(\mu + \Sigma_{1/2}z)} = \sum_{k=1}^K \frac{\Sigma_{1/2} g_{jk}(\mu + \Sigma_{1/2}z)}{g_j(\mu + \Sigma_{1/2}z)}$$

où C_l représente le dénombrement pour le l -ième petit domaine et \sum_p représente la sommation des huit petits domaines de la province p . Une erreur quadratique moyenne pour le facteur de redressement provincial, soit conditionnelle, soit inconditionnelle, peut être calculée à l'aide d'une matrice mpe d'erreurs du produit moyen sous la forme

$$\text{mse}(\hat{F}^p) = \frac{1}{I} \sum_d^d \sum_d^d C_l' \text{mpe}(\hat{F}_l', \hat{F}_l')$$

Des erreurs quadratiques moyennes conditionnelles sont obtenues à l'aide de la formule (2) pour mpe. Lahiri et Rao (1995) ont présenté une formule pour les termes hors diagonale de la matrice de l'erreur inconditionnelle du produit moyen dont la diagonale est donnée par les erreurs quadratiques moyennes de Prasad et Rao (1990).

Tableau 4

Estimations directes (F^d) et empiriques de Bayes (F^b) des facteurs de correction provinciaux avec leurs efficacités conditionnelles (eff^{pc}) et inconditionnelles (eff^{ppr}). Une efficacité conditionnelle est ∞ lorsque l'estimateur de l'erreur quadratique moyenne conditionnelle est nul

PROVINCE	F^d	F^b	eff^{pc}	eff^{ppr}
Terre-Neuve	1,203	1,0176	6,49	2,94
Ile-du-Prince-Édouard	1,0094	1,0153	1,03	4,52
Nouvelle-Écosse	1,0193	1,0171	25,3	2,59
Nouveau-Brunswick	1,0335	1,0367	0,67	1,11
Québec	1,0268	1,0262	1,12	0,93
Manitoba	1,019	1,0176	∞	2,46
Saskatchewan	1,0183	1,0166	∞	2,54
Alberta	1,0204	1,0187	7,37	1,98
Colombie-Britannique	1,0281	1,0293	1,09	1,03
Yukon	1,0396	1,040	1,41	1,17
Territoires du N.-O.	1,0575	1,055	1,40	1,32

Le tableau 4 présente des estimations agrégées directes et empiriques de Bayes avec deux efficacités. Les estimations empiriques de Bayes retiennent une bonne partie des différences interprovinciales. C'est là une indication que les variables explicatives du modèle de lissage ont capté la plupart des différences entre les taux de sous-dénombrement provinciaux. Une exception notable est l'Ile-du-Prince-Édouard, dont le faible facteur de correction n'est pas élucidé par les variables explicatives. Il s'agit de la seule province pour laquelle les deux efficacités sont appréciablement différentes. Les efficacités conditionnelles sont plus instables que les efficacités de Prasad et Rao. Sauf à l'Ile-du-Prince-Édouard, les deux décrivent un phénomène semblable: au Nouveau-Brunswick, au Québec, en Ontario et en Colombie-Britannique, les estimations empiriques de Bayes agrégées ne représentent pas une amélioration appréciable relativement aux estimateurs directs.

Il est également possible de calculer la seconde dérivée partielle de $g_i^*(r)$; elle porte le même signe que $r_i - r_N$. Ainsi, l'asymétrie positive du taux de sous-dénombrement, qui est probable lors de l'estimation d'événements rares comme le fait d'être manqué dans un recensement, entraîne un accroissement de l'erreur quadratique moyenne conditionnelle dans les provinces qui affichent un sous-dénombrement supérieur au taux national.

Pour 1991, $\hat{\alpha} = 0,874$ et le taux de sous-dénombrement national est $r_N = 2,872\%$. Le tableau 2 présente les estimations du sous-dénombrement provincial r_i^c avec leurs efficacités $eff_i^c = \sigma_{ii}^2 / mse(r_i^c)$, où $mse(r_i^c)$ est calculée suivant la définition donnée à la section 2, en fonction de la correction proposée à la section 3.2 pour rendre compte des variances estimées. L'estimateur composé représente une amélioration relativement aux estimateurs directs dans tous les cas sauf trois qui correspondent aux provinces affichant les taux de sous-dénombrement les plus extrêmes.

Le tableau 2 fournit également l'estimateur empirique de Bayes r_i^B calculé à l'aide d'un modèle de lissage de l'emplacement. Dans le cadre du modèle (M), le véritable taux de sous-dénombrement θ_i est supposé comporter une distribution sous forme de $N(\beta, \sigma_\theta^2)$. Les estimations paramétriques sont $\hat{\sigma}_\theta^2 = 1,45 \times 10^{-4}$ et $\hat{\beta} = 2,61\%$. Deux efficacités sont présentées relativement à des estimateurs directs, soit d'une part eff_i^B qui est calculée à l'aide de l'estimateur de l'erreur quadratique moyenne conditionnelle pour r_i^B , y compris la correction décrite à la section 3.2 pour tenir compte des variances estimées, et d'autre part eff_i^{pp} qui est calculée à l'aide de l'estimateur inconditionnel de Prasad et Rao. Le taux important de sous-dénombrement pour les Territoires du Nord-Ouest est responsable de l'estimation élevée pour $\hat{\sigma}_\theta^2$; c'est pourquoi les estimateurs empiriques de Bayes r_i^B se rapprochent beaucoup plus des estimateurs directs r_i^c que les estimateurs composites. Une reprise de l'analyse sans les Territoires du

Tableau 2
Deux estimateurs du sous-dénombrement provincial et leurs efficacités

Province	r_i^c	CV	r_i^c	eff_i^c	r_i^B	eff_i^B	eff_i^{pp}
Terre-Neuve	2,06	1,994	15,96	2,105	1,12	2,038	1,04
Ile-du-Prince-Édouard	0,47	0,931	30	1,176	0,65	1,025	0,93
Nouvelle-Écosse	3,26	1,889	20,05	2,013	1,11	1,959	1,06
Nouveau-Brunswick	2,66	3,245	13,73	3,198	1,29	3,162	1,09
Québec	25,19	2,605	8,35	2,639	1,16	2,605	1,02
Ontario	37,24	3,641	8,46	3,544	0,87	3,572	1,04
Manitoba	3,96	1,86	20,83	1,987	1,1	1,936	1,06
Saskatchewan	3,58	1,798	18,87	1,933	1,04	1,863	1,05
Alberta	9,24	1,995	14,57	2,106	1,01	2,032	1,06
Colombie-Britannique	12,01	2,733	9,86	2,751	1,26	2,727	1,07
Yukon	0,1	3,83	15,99	3,709	1,27	3,56	1,05
Territoires du Nord-Ouest	0,22	5,439	11,28	5,116	0,96	4,813	0,49
							1,18

Nord-Ouest et le Yukon modifie énormément les estimations empiriques de Bayes. Dans le tableau 2, l'estimateur composite fonctionne mieux que l'estimateur empirique de Bayes; il entraîne un accroissement de l'efficacité conditionnelle supérieur à 10% dans sept des 12 provinces. Trois efficacités sont inférieures à 1; la discussion de la section 4.2 laisse entendre que les efficacités inférieures à 1 sont inévitables. La précision relative médiocre des $\hat{\sigma}_\theta^2$ (leur estimation se fait selon quatre degrés de liberté seulement) diminue les efficacités conditionnelles des estimateurs empiriques de Bayes. Elle exerce une influence moindre sur l'estimateur composite puisque le même paramètre de réduction est utilisé pour toutes les provinces. Les efficacités conditionnelles captent les performances médiocres de r_i^c et r_i^B dans les provinces ayant les taux de sous-dénombrement les plus extrêmes. Or les efficacités de Prasad et Rao les manquent. Elles soulignent les améliorations que le lissage entraîne pour les deux territoires où les taux de sous-dénombrement sont très variables. Les efficacités de Prasad et Rao ont un sens uniquement si l'on accepte l'hypothèse de provinces échangeables qui sous-tend le modèle de lissage. Cela est douteux puisque le sous-dénombrement tend à être plus élevé dans les provinces à forte population urbaine que dans les petites régions rurales.

5.2 Estimations intra-provinciales

Dick (1995) a examiné l'estimation des facteurs de redressement pour le sous-dénombrement du recensement de 1991 (catégories âge x sexe pour chaque province). Le facteur de redressement pour un petit domaine se définit comme $F=1+$ (sous-dénombrement estimatif)/(dénombrement). Pour quatre catégories d'âge, 0-19, 20-29, 30-44, 45+ et deux sexes, il existe 96 petits domaines. Les variables explicatives sont des interactions entre les variables des indicateurs pour les 12 provinces, les quatre groupes d'âge

personnes comptées deux fois ou comptées de façon erronée dans le recensement, et la Contre-vérification des dossiers (Burgess, 1988) pour les personnes manquées dans le Recensement. Le fait de combiner ces chiffres donne une estimation du sous-dénombrement du Recensement. La présente section examine divers estimateurs du sous-dénombrement d'un recensement.

5.1 Estimations provinciales

Les taux de sous-dénombrement de 1991 pour les dix provinces du Canada et ses deux territoires et leurs coefficients de variation sont présentés dans le tableau 2. La proportion p_i de la population habitant chacune des provinces (le mot province désigne dans la présente section les dix provinces du Canada et les deux territoires) y est également indiquée. Les coefficients de variation (c.v.) du tableau 2 ont été calculés à partir de variances estimées à l'aide de cinq groupes aléatoires. Ainsi, les variances de l'échantillonnage peuvent être considérées comme comportant une distribution χ^2_4 . Partout dans la présente section, nous supposons que les estimations du sous-dénombrement provincial et leurs variances sont indépendantes.

Royce (1992) a proposé plusieurs estimateurs du sous-dénombrement provincial. Rivest (1995) a proposé un estimateur composite qui réduit le taux de sous-dénombrement provincial vers le taux national. Il est donné par

$$r_i^c = \hat{\alpha} r_i + (1 - \hat{\alpha}) r_N,$$

où $r_N = \sum p_i r_i$ est le taux de sous-dénombrement national et le paramètre de réduction $\hat{\alpha}$ est donné par:

$$\hat{\alpha} = \frac{\sum p_i' (1 - p_i') \sigma_i'^2 + \sum p_i' r_i'^2 - r_N'^2}{\sum p_i' r_i'^2 - r_N'^2}.$$

C'est là la valeur de α qui est optimale pour les fonctions de perte de l'estimation des totaux provinciaux et de la part provinciale de la population (on trouvera des détails dans Royce (1992) et dans Rivest (1995)). On a $r_i^c = r_i + g_i(r_i)$, où

$$g_i(r_i) = - \frac{\sum p_i' (1 - p_i') \sigma_i'^2}{(r_i' - r_N') \left(\sum p_i' (1 - p_i') \sigma_i'^2 + \sum p_i' r_i'^2 - r_N'^2 \right)}.$$

On peut facilement calculer une expression analytique pour l'estimateur de l'erreur quadratique moyenne conditionnelle en notant que

$$\frac{\partial g_i(r_i)}{\partial r_i} = 2p_i' (r_i' - r_N')^{-2} \left[\sum p_i' (1 - p_i') \sigma_i'^2 + \sum p_i' r_i'^2 - r_N'^2 \right] - (1 - p_i') (1 - \hat{\alpha}).$$

tionnelle des estimateurs empiriques régionaux de Bayes

satisfait

$$E_M[E_S(\hat{q}_i) | \text{mse}_c(\hat{q}_i)] = \text{MSE}(\hat{q}_i),$$

où $\text{MSE}(\hat{q}_i)$ est la variance inconditionnelle liée au modèle

prédit.

La proposition 5 montre que $\text{mse}_c(\hat{q}_i)$ peut être considérée comme une étape intermédiaire de l'évaluation de l'erreur quadratique moyenne inconditionnelle de \hat{q}_i . Considérons par exemple le calcul de l'approximation

$O(1/n)$ de Prasad et Rao (1990) pour $\text{MSE}(\hat{q}_i)$,

$$\text{MSE}_{\text{PR}}(\hat{q}_i) = \frac{\sigma''^2 \sigma_i'^2}{\sigma''^2 + \sigma_i'^2} + \frac{\sigma''^2 x_i' A^{-1} x_i'}{\sigma''^2 + \sigma_i'^2} + \frac{\sigma''^2 \text{Var}(\hat{q}_i^2)}{(\sigma''^2 + \sigma_i'^2)^3}.$$

La dérivation standard, examinée à la section 3.2 de Singh, Stukel, et Pfeffermann (1998), se fonde sur Kackar et Harville (1984). Une autre dérivation, présentée dans Belmonte (1998, 1999), consiste à prendre l'espérance de $\text{mse}_c(\hat{q}_i)$, obtenue à l'aide de (8), relativement à la distribution marginale des y_i' , qui sont des écarts typiques de variable indépendante $N(x_i' \beta, \sigma''^2 + \sigma_i'^2)$, puis à ne retenir que les termes d'ordre supérieur.

La proposition 5 s'applique lorsque les estimateurs régionaux sont évalués, ou lorsque les corrections suggérées à la section 3 sont mises en œuvre. Ce sont là des cas pour lesquels il n'existe pas de formule analytique pour les variances liées au modèle prédit. La proposition 4 indique une méthode simple de construction des estimations de Monte Carlo inconditionnelles. Il suffit de produire un grand nombre de répétitions de $\{y_i', i = 1, \dots, n\}$ où y_i' suit une $N(x_i' \beta, \sigma''^2 + \sigma_i'^2)$ et de calculer $\text{mse}_c(\hat{q}_i)$ pour chacune. L'établissement d'une moyenne des $\text{mse}_c(\hat{q}_i)$ fournit une variance inconditionnelle liée au modèle prédit qui est identique à l'EQM de la proposition 4 évaluée pour des estimations β'' , $\sigma_i'^2$ des paramètres inconnus. Malheureusement, cette estimation est biaisée (il s'agit d'une estimation d'ordre 1 dans la terminologie de Singh, Stukel et Pfeffermann (1998)). Pour ce qui est de l'estimateur empirique de Bayes donné en (7), d'après (9) le biais de l'estimation de Monte Carlo calculée à partir de la proposition 4 est $-\sigma''^2 \text{Var}(\hat{q}_i^2) / (\sigma''^2 + \sigma_i'^2)^3$. Il va falloir poursuivre les travaux pour construire, à partir de la proposition 4, des estimateurs inconditionnels sans biais de la variance liée au modèle prédit.

5. ESTIMATION DU SOUS-DÉNOMBREMENT DU RECENSEMENT CANADIEN DE 1991

En 1991, le sous-dénombrement du Recensement du Canada a été estimé à l'aide de deux enquêtes, l'étude du

surdénombrement qui a permis d'estimer le nombre de

absolus, définis comme $|\sum^* (mse_i(\hat{\mu}_i) - MSE_i)| / (mMSE_i)$ de même que de leurs coefficients de variation, identiques à $(\sum^* (mse_i(\hat{\mu}_i) - MSE_i)^2 / m) / MSE_i$.

Tableau 1

Efficacité relative des estimateurs empiriques de Bayes (ER), biais relatif absolu (BR) et coefficient de variation (CV) de deux estimateurs de l'EQM ($n = 30$). Tous les résultats sont exprimés en pourcentage

$\sum (\mu_i - \hat{\mu})^2 / 29$	ER %	BR %	CV %	BR %	CV %	BR %	CV %
1,3	21,2	1	4,7	97	51	43	31
2,53	14,9	2	3,0	37	32	20	23
3,7	12,9	2	2,0	21	24	21	20
4,24	12,5	2	1,9	19	20	22	22
4,93	12,2	1	1,7	15	18	17	13
	moyenne	1,33	1	17	13	17	13
	médiane	1,33	1	17	13	17	13

Dans certaines situations, comme celle dont il est

question à la section 5.1, la réduction est faible et le recours

à un estimateur conditionnel de l'erreur quadratique moyenne est approprié. L'efficacité conditionnelle de $\hat{\mu}_i$ relativement à l'estimateur direct y_i est donnée par

$\sigma'' / mse_c(\hat{\mu}_i)$. Elle est supérieure à 1 à la condition que

$(y_i - x_i' \beta) / (\sigma'' + \sigma_v^2) < 2$. Si l'on suppose que le modèle de

lissage s'applique, on peut s'attendre à des efficacités

conditionnelles inférieures à 1 pour 16% environ

$(P[N(0,1) > 2])$ des estimateurs régionaux. Ce pour-

centage devrait être plus élevé lorsque le modèle de lissage

est déficient. Des efficacités conditionnelles inférieures

à 1 surviennent dans des petites régions comportant de

grandes valeurs résiduelles. Par contre, les efficacités

inconditionnelles, calculées à l'aide de la variance posté-

rieure, sont dans un tel cas inférieures à 1 pour toutes les

petites régions. Cela indique qu'il est pratiquement

impossible que toutes les efficacités conditionnelles soient

inférieures à 1; Rao et Shinozaki (1978) l'avaient déjà

remarqué pour les estimateurs de James-Stein.

Plusieurs des observations formulées dans la situation

connus s'appliquent lorsque des paramètres sont estimés.

L'erreur quadratique moyenne est l'estimateur conditionnel de

Rao (1990),

$$mse_{PR}(\hat{\mu}_i) = \frac{\sigma''^2 \sigma_v^2}{\sigma''^2 + \sigma_v^2} + \frac{\sigma''^2 x_i' A_i^{-1} x_i}{2(\sigma''^2 + \sigma_v^2)^2} + \frac{\sigma''^2 \widehat{Var}(\hat{\sigma}_v^2)}{(\sigma''^2 + \sigma_v^2)^3}, \quad (9)$$

où $\widehat{Var}(\hat{\sigma}_v^2) = 2 \sum (\hat{\sigma}_v'' + \hat{\sigma}_v^2)^2 / n^2$. Afin de vérifier dans

quelle mesure la proposition 4 s'applique lorsque des

paramètres sont estimés, une petite étude de Monte Carlo a

été menée en fonction de l'étude de simulation de la

stratégie ii) de Prasad et Rao (1999). Dans toutes les

simulations, $n = 30$ et $\sigma'' = 1$, pour $i = 1, \dots, n$. Le modèle

de lissage (6) était $\mu_i = \mu + v_i$ et diverses valeurs de σ_v^2 ont

été utilisées. Les résultats indiqués au tableau 1 se fondent

sur $m = 5000$ répétitions de Monte Carlo.

Les simulations ont utilisé cinq ensembles de valeurs μ_i

dont la variance est indiquée au tableau 1. Pour chaque

ensemble, y_i a été simulé de façon répétée sous forme de

variable aléatoire $N(\mu_i, 1)$, $i = 1, \dots, n$. L'estimation empi-

rique de Bayes $\hat{\mu}_i$ a été calculée pour chaque petite région,

et l'erreur quadratique moyenne pour la petite région i a été

calculée: $MSE_i = \sum^* (\hat{\mu}_i - \mu_i)^2 / m$ où \sum^* désigne la

somme des m répétitions de Monte Carlo. L'efficacité de

l'estimateur empirique de Bayes pour la petite région i est

$1 / MSE_i$. La moyenne et la médiane de $n = 30$ efficacités

régionales sont données au tableau 1. Les deux erreurs

quadratiques moyennes (conditionnelle et inconditionnelle)

ont été calculées pour chaque petite région dans les m

répétitions de Monte Carlo; selon (9), $mse_{PR}(\hat{\mu}_i) =$

$(\hat{\sigma}_v'' + 5/n) / (1 + \hat{\sigma}_v^2)$ pour chaque petite région. Le tableau

1 présente la moyenne et la médiane de leurs biais relatifs

Comme il a été montré à la section 2, $mse_c(\hat{\mu}_i)$ est sans

4.3 Erreur quadratique moyenne conditionnelle et variance liée au modèle prédictif

La présente section examine le rapport entre l'erreur

quadratique moyenne conditionnelle proposée dans le

présent exposé et la variance liée au modèle prédictif qui est

une mesure inconditionnelle de l'exactitude. Si l'on utilise

la rotation de (6), la variance liée au modèle prédictif est

$MSE_B(\hat{\mu}_i) = E_M[E_S\{(\hat{\mu}_i - x_i' \beta - v_i)^2\}]$. Compte tenu de la

construction présentée à la section 2, on a

$$E_S\{mse_c(\hat{\mu}_i)\} = E_S\{(\hat{\mu}_i - x_i' \beta - v_i)^2\}.$$

La meilleure façon d'examiner les propriétés de l'estimateur de l'erreur quadratique moyenne conditionnelle est de considérer la situation simple dans laquelle tous les paramètres du modèle de lissage sont supposés connus. Dans une telle situation, $\partial g_i(V) / \partial y_i = -\sigma'' / (\sigma'' + \sigma_v^2)$, et l'estimateur de l'erreur quadratique moyenne conditionnelle est identique à $\text{mse}_c^*(\hat{\mu}_i) = \max\{\text{mse}_c(\hat{\mu}_i), 0\}$ où

$$\text{mse}_c(\hat{\mu}_i) = \frac{\sigma'' \sigma_v^2}{\sigma'' + \sigma_v^2} + \left(\frac{\sigma'' + \sigma_v^2}{\sigma''} \right) \left\{ (y_i - x_i' \beta)^2 - \sigma'' - \sigma_v^2 \right\}.$$

L'alternative fondée sur un modèle pour cet estimateur est $E_M[E_S\{\text{mse}_c(\hat{\mu}_i)\}]$. Cet estimateur, qui est un cas spécial de l'estimateur de Prasad et Rao (1990), est noté $\text{mse}_{\text{pr}}(\hat{\mu}_i)$. L'estimateur $\text{mse}_c^*(\hat{\mu}_i)$ est très variable lorsque σ_v^2 est faible. En effet, lorsque σ_v^2 se rapproche de 0, la moitié environ des estimations de l'erreur quadratique moyenne conditionnelle sont nulles. Pour mieux comparer les deux estimateurs de l'erreur quadratique moyenne, conditionnel et inconditionnel, de $E_S\{\text{mse}_c(\hat{\mu}_i)\}$.

La proposition suivante compare les fluctuations moyennes des estimateurs, conditionnel et inconditionnel, est

$$E_S\{\text{mse}_c(\hat{\mu}_i)\} = \frac{\sigma'' \sigma_v^2}{\sigma'' + \sigma_v^2} + \left(\frac{\sigma'' + \sigma_v^2}{\sigma''} \right) \left\{ (\mu_i - x_i' \beta)^2 - \sigma_v^2 \right\}.$$

PROPOSITION 4: Lorsque $\sigma'' = \sigma_v^2$, pour $i = 1, \dots, n$ et lorsque les moyennes régionales sont des μ_i tirées à l'aide de (6), l'efficacité de la variance postérieure relativement à l'estimateur de l'erreur quadratique moyenne conditionnelle est

$$\frac{E_M[\sum \text{MSE}_S\{\text{mse}_c(\hat{\mu}_i)\} / n]}{E_M[\sum \text{MSE}_{\text{pr}}(\hat{\mu}_i) / n]} = \frac{\sigma_v^4}{\sigma_v^4 + 2\sigma_v^2}.$$

où $\text{MSE}_S(\cdot)$ désigne une erreur quadratique moyenne prise relativement à la distribution des y_i qui sont des variables aléatoires $N(\mu_i, \sigma_v^2)$ indépendantes. L'efficacité ci-dessus est supérieure à 1 à la condition que $\sigma_v^2 / \sigma'' < 2,41$. La proposition 4 indique que, en présence d'une forte réduction, l'estimateur inconditionnel de l'erreur quadratique moyenne est un meilleur estimateur de l'erreur quadratique moyenne conditionnelle que ne l'est l'estimateur conditionnel. Ce résultat surprenant est attribuable à la variance élevée de l'estimateur conditionnel; c'est un estimateur médioere lorsque la réduction est considérable.

On peut obtenir une forme explicite de (3) à partir de la formule ci-dessous pour la dérivée des fonctions g_i pour les estimateurs empiriques de Bayes,

$$\frac{\partial g_i(V)}{\partial y_i} = \frac{\partial g_i(V)}{\partial \sigma_v^2} \frac{\partial \sigma_v^2}{\partial y_i} - \frac{\sigma''}{\sigma'' + \sigma_v^2} \left\{ 1 - \frac{x_i' \lambda^{-1} x_i}{x_i' \lambda^{-1} x_i + \sigma''} \right\}, \quad (8)$$

Les dérivées partielles qui figurent en (8) peuvent être évaluées à l'aide de méthodes standard. Elles sont données par

$$\frac{\partial \sigma_v^2}{\partial y_i} = \frac{(n-d)}{2} (y_i - x_i' \beta),$$

et par

$$\frac{\partial g_i(V)}{\partial \sigma_v^2} = \frac{\sigma''}{(\sigma_v^2 + \sigma'')^2} (y_i - x_i' \beta'') + \frac{\sigma''}{(\sigma_v^2 + \sigma'')^2} x_i' \beta''.$$

où

$$\frac{\partial \beta''}{\partial \sigma_v^2} = -\lambda^{-1} \sum_n x_i (y_i - x_i' \beta'') \frac{(\sigma_v^2 + \sigma'')^{-2}}{(\sigma_v^2 + \sigma'')^2}.$$

Selon (8), on a une expression de forme analytique pour $\text{mse}_c(\hat{\mu}_i)$. Cette statistique est un estimateur de l'erreur quadratique moyenne pour l'estimateur empirique de Bayes pour la i -ième petite région relativement au plan d'échantillonnage uniquement. Elle est valable pour toute taille d'échantillon n ; elle se fonde sur la seule hypothèse d'une distribution normale des estimateurs directs y_i . Lorsque $\sigma_v^2 = 0$, $\hat{\mu}_i = x_i' \beta''$ et les dérivées en (8) se simplifient de manière appréciable. Puisque $\partial \sigma_v^2 / \partial y_i = 0$, on a

$$v(\hat{\mu}_i) = \sigma''(1 - 2\alpha) + (\hat{\alpha}_i y_i - \hat{\gamma}_i)^2$$

Cela est égal à (3) lorsque $(d/dy_i)\hat{\alpha}_i$ et $(d/dy_i)\hat{\gamma}_i$ sont tous deux nuls. Ainsi, l'estimateur de Kott (1989) pour l'erreur quadratique moyenne conditionnelle ne tient pas compte de l'estimation pour les composantes de la variance. C'est peut-être ce qui explique le biais manifesté par cet estimateur dans les simulations décrites par Prasad et Rao (1999). Les estimations mse_c et mpe_c se laissent évaluer numériquement si l'on prend

$$\frac{\partial g_i(y)}{\partial g_i(y)} = \frac{2\epsilon}{-g_i(y_1, \dots, y_{j-1}, y_j - \epsilon, y_{j+1}, \dots, y_n) - g_i(y_1, \dots, y_{j-1}, y_j + \epsilon, y_{j+1}, \dots, y_n)}$$

où ϵ est un petit nombre positif. Ainsi mse_c et mpe_c se laissent calculer dans toutes les situations, même lorsque g n'a aucune forme explicite.

Afin d'illustrer la souplesse de l'estimateur conditionnel, considérons $\hat{\mu}^* = \hat{\mu}(\sum y_i) / (\sum \hat{\mu}_i)$, un estimateur évalué de façon à s'accorder avec l'estimateur direct pour le total y .

On a $\hat{\mu}^* = y + g^*(y)$ où

$$g^*(y) = \frac{\sum y_i g(y)}{\sum y_i} + \left(\frac{\sum \hat{\mu}_i}{\sum y_i} - 1 \right) y.$$

Il peut être difficile de calculer une formule analytique pour $mpe_c(\hat{\mu}^*)$, mais cette expression est facile à évaluer à l'aide de données numériques. On trouvera ci-dessous des modifications de l'estimateur conditionnel permettant d'expliquer la non-normalité des y_i et les variances estimées σ'' .

3. ANALYSE DE SENSIBILITÉ

Dans plusieurs enquêtes, surtout celles du secteur des entreprises, les variables à l'étude sont asymétriques. Il est possible qu'une partie de cette asymétrie subsiste dans les estimateurs directs y_i . La présente section propose une correction de l'erreur quadratique moyenne conditionnelle afin d'expliquer l'asymétrie de la distribution de y tout en décrivant des façons d'expliquer l'estimation des variances σ'' dans les calculs de l'erreur quadratique moyenne.

Concrètement, on estime les variances σ'' . Plusieurs auteurs (Dick 1995; Hogan 1992) assurent le lissage des variances avant le calcul des estimations régionales. Ils considèrent alors les variances lissées comme les véritables variances dans le calcul des petites régions. La section 3.2 trouve également une description de situations dans lesquelles les variances d'échantillonnage sont estimées à l'aide de groupes aléatoires (Wolter 1985, ch.2). Suivant cette méthode, on réalise un certain nombre, k par exemple, de répétitions du plan d'enquête. Cela donne, pour chaque

i , k estimations de μ_i ; $\hat{\sigma}_i''$ est alors égal à la variance d'échantillonnage de ces k estimations divisée par k . Si l'on suppose que ces k estimations ont une distribution normale, il est possible de considérer que, convenablement normalisée, la distribution de $\hat{\sigma}_i''$ est une distribution chi carré à $k - 1$ degrés de liberté. Une erreur quadratique moyenne conditionnelle, corrigée en fonction de variances estimées à l'aide de groupes aléatoires, est proposée dans la présente section. Afin que la discussion reste simple, nous supposons ici que Σ est une matrice diagonale; autrement dit, les y_i sont supposés représenter des variables aléatoires indépendantes.

3.1 Non-normalité de la distribution des y_i

Dans plusieurs applications de l'estimation régionale, la distribution des y_i n'est pas exactement normale. Nous posons une simple correction de (3) pour traiter l'asymétrie de la distribution des y_i .

Supposons que l'asymétrie des y_i , $p_i = E\{y_i - \mu_i\}^3 / \sigma_i''^3$ est faible et non nulle. Une série d'Edgeworth d'ordre un pour la distribution des y_i est donnée (voir par exemple Reid 1991) par:

$$f(t) = \frac{\exp\{-t - \mu_i\}^2 / (2\sigma_i'')\}}{\sqrt{2\sigma_i''\pi}} \times \left[1 + \frac{p_i}{6} \left(\frac{t - \mu_i}{\sigma_i''} \right)^3 - 3 \left(\frac{t - \mu_i}{\sigma_i''} \right) \left(\frac{\sqrt{\sigma_i''}}{\sigma_i''} \right) \right]$$

Une telle expansion sert à corriger l'asymétrie des estimateurs directs (Barndorff-Nielsen et Cox 1989, remarque 2, p. 92). On utilise des expansions comportant des termes additionnels pour la correction de l'asymétrie aussi bien que de l'aplatissement; il n'en est pas question dans la présente section. L'évaluation de $E\{y_i - \mu_i\}^3 g_i(y)$ dans le cadre de f_i nécessite à la construction de l'estimateur de l'erreur quadratique moyenne conditionnelle, est donnée dans la proposition 2.

PROPOSITION 2: Lorsque les y_i sont distribués en conformité avec $f_i(t)$, et que p_i tend vers 0,

$$E_S\{y_i - \mu_i\} g_i(y) = \left\{ \frac{\sigma_i'' E_S\left\{ \frac{\partial y_i}{\partial g_i(y)} \right\}}{\sigma_i'' p_i} + \frac{2}{\sigma_i'' p_i} E_S\left\{ \frac{\partial y_i^2}{\partial g_i(y)} \right\} \right\} + O(p_i).$$

Un estimateur de l'erreur quadratique moyenne corrigé pour l'asymétrie est donc donné par $mse_c^*(\hat{\mu}_i) = \max\{0, mse_c(\hat{\mu}_i)\}$ où

$$mse_c(\hat{\mu}_i) = \sigma'' + 2\sigma'' \frac{\partial y_i}{\partial g_i(y)} + \sigma_i'' p_i \frac{\partial^2 g_i(y)}{\partial y_i^2} + g_i(y)^2.$$

Concrètement, il peut être difficile de trouver des coefficients d'asymétrie individuels p_i pour chaque i . Une

2. UN ESTIMATEUR DE L'ERREUR QUADRATIQUE MOYENNE CONDITIONNELLE

L'espérance des écarts typiques d'une variable normale. La section 3 propose des modifications de l'estimateur conditionnel pour tenir compte de l'asymétrie de la distribution des estimateurs directs et des variances estimées. La section 4 décrit l'application du nouvel estimateur aux estimateurs empiriques de Bayes. Il est question notamment de son rapport avec la variance liée au modèle prédictif (Prasad et Rao 1990). Des exemples sont présentés à la section 5.

Supposons qu'il existe n petites régions; soit $\mu = (\mu_1, \dots, \mu_n)'$ les caractéristiques inconnues de la population pour ces petites régions. Les estimations d'enquête

directes pour les n petites régions sont $y = (y_1, \dots, y_n)'$ où la distribution de y est $N_n(\mu, \Sigma)$, une distribution normale n -variée comportant un vecteur moyen μ et une matrice de variance-covariance connue Σ . Comme il a été montré par Ghosh et Rao (1994), l'hypothèse de normalité demeure probablement valable pour de nombreuses enquêtes puisque les estimations d'enquête directes sont normalement des fonctions de sommes de variables. La matrice $n \times n$ Σ est une mesure de la précision pour y fondée sur le plan de sondage. Pour le moment, cette matrice est supposée connue. Cette hypothèse est associée à la section 3.2. L'incertitude liée à y provient de la sélection aléatoire des unités d'échantillonnage. L'indice S , pour le plan de sondage, désigne l'espérance prise relativement à la distribution de y .

Dans une application typique des techniques régionales, on a

$$y_i = \frac{\sum_j w_{ij} y_j}{\sum_j w_{ij}}$$

où y_j est la valeur y de la j -ième unité d'échantillonnage de la petite région i , w_{ij} est son poids d'échantillonnage et la somme englobe toutes les unités d'échantillonnage de la petite région i . Dans plusieurs cas, la matrice de variance-covariance Σ est diagonale; son terme (i, i) est $\sigma_{ii}^2 = \text{Var}_S(y_i)$; lorsqu'ils ne sont pas nuls, les éléments hors diagonale de Σ sont notés σ_{ij} , $i, j = 1, \dots, n$.

On a proposé plusieurs méthodes d'amélioration de l'exactitude des estimateurs d'enquête directs. Elles comportent une réduction de y_i vers un estimateur indirect quelconque de μ_i . Les estimateurs qui en résultent s'écrivent sous la forme

$$\hat{\mu}_i = y_i' + g_i'(y^1, \dots, y^n), \quad i = 1, \dots, n \quad (1)$$

où les g_i' sont des fonctions qui dépendent de la stratégie de réduction. Sous forme vectorielle, (1) s'écrit $\hat{\mu} = y + g(y)$ où g , dont le i -ième élément est égal à g_i' , est une fonction définie de R^n à R^n . Nous supposons que, pour chaque i , la

$$E_S\{(\hat{\mu} - \mu)(\hat{\mu} - \mu)'\} = \Sigma + E_S\{(y - \mu)(y - \mu)'\} + E_S\{g(y)(g(y))'\}$$

dérivée partielle de droite et la dérivée partielle de gauche de g_i' relativement à y_j existent pour tout y en R^n . Lorsqu'elles sont égales, $\partial g_i(y)/\partial y_j$ désigne la valeur commune; si elles sont différentes, $\partial g_i(y)/\partial y_j$ est la moyenne des deux valeurs. La composante de $g(y)$ et ses dérivées partielles sont supposées comporter des variances finies. Une évaluation conditionnelle de la précision de $\hat{\mu}$ comme estimateur de μ est donnée par la matrice des erreurs du produit moyen qui est donnée par

PROPOSITION 1: Soit y un vecteur aléatoire $N_n(\mu, \Sigma)$; on a alors:

$$E_S\{(y - \mu)(g(y))'\} = \Sigma E_S\{\nabla g(y)\},$$

où $\nabla g(y)$ est une matrice $n \times n$ dont le (i, j) -ième élément est donné par $g_i'(y) = \partial g_i(y)/\partial y_j$.

En conformité avec la proposition 1, $\Sigma \nabla g(y)$ est un estimateur sans biais pour $E_S\{(y - \mu)(g(y))'\}$. Ainsi, l'estimateur conditionnel (l'indice «c» signifie conditionnel) pour la matrice des erreurs du produit moyen est donné par

$$\text{mpe}_c(\hat{\mu}) = \Sigma + \Sigma \nabla g(y) + \nabla g(y)' \Sigma + g(y)(g(y))' \quad (2)$$

Les termes diagonaux de (2) peuvent être négatifs. Puisqu'ils servent à estimer les erreurs quadratiques moyennes d'un meilleur estimateur pour l'erreur quadratique moyenne de $\hat{\mu}_i$ est

$$\text{mse}_c(\hat{\mu}_i) = \max \left(0, \sigma_{ii}'' + \sum_j \sigma_{ij}^2 \{g_j'(y) + g_j'(y)\} + g_i(y)^2 \right).$$

Il généralise un estimateur proposé par Bildeau et Srivastava (1987) pour l'estimateur de James-Stein et par Robert (1992, p. 292) pour les estimateurs empiriques de Bayes. Lorsque les y_i sont indépendants, suivant $\sigma_{ij} = 0$ lorsque $i \neq j$, on a

$$\text{mse}_c(\hat{\mu}_i) = \sigma_{ii}'' + 2\sigma_{ii}'' \frac{\partial g_i(y)}{\partial y_i} + g_i(y)^2 \quad (3)$$

et $\text{mse}_c(\hat{\mu}_i) = \max\{\text{mse}_c(\hat{\mu}_i), 0\}$. L'estimateur régional de Kott (1989) comporte $g_i'(y) = \hat{g}_i'(y_i - y_i)$, où \hat{y}_i est une mesure de l'emplacemement des y et \hat{g}_i est un paramètre de lissage. Ces deux statistiques comprennent des estimations de la variance calculées au niveau «unité», c'est-à-dire à l'aide des y_j . L'erreur quadratique moyenne conditionnelle de Kott (1989) est

Une erreur quadratique moyenne conditionnelle des estimateurs régionaux

LOUIS-PAUL RIVEST et EVE BELMONTÉ¹

RÉSUMÉ

Les auteurs proposent l'estimation de l'erreur quadratique moyenne conditionnelle des estimateurs régionaux comme moyen d'en évaluer la précision. Cette erreur quadratique moyenne est conditionnelle en ce sens qu'elle mesure la variabilité relativement au plan d'échantillonnage pour une réalisation particulière du modèle de lissage qui sous-tend les estimateurs régionaux. Il est facile de construire un estimateur sans biais pour l'erreur quadratique moyenne conditionnelle à l'aide du lemme de Stein pour l'espérance de variables aléatoires normales. On peut calculer cet estimateur pour toute stratégie de réduction; les auteurs considèrent les estimateurs composites et empiriques de Bayes. L'estimateur peut être calculé lorsque les estimateurs régionaux sont évalués de façon à correspondre à des estimateurs directs pour un niveau d'agrégation élevé. L'estimateur peut tenir compte de l'asymétrie des données et des variances estimées. L'estimateur de l'erreur quadratique moyenne conditionnelle ne dépend d'aucun modèle de lissage. Cette propriété est assurée moyennant une variance élevée; le nouvel estimateur est instable en présence de réduction poussée. Dans de telles situations, il fournit néanmoins de précieux renseignements diagnostiques au sujet du modèle de réduction. On peut également le considérer comme un élément de l'erreur quadratique moyenne inconditionnelle comme celui de Prasad et Rao (1990). Des exemples portant sur l'estimation du sous-dénombrement du Recensement du Canada illustrent l'application de ce nouvel estimateur.

MOTS CLÉS: Sous-dénombrement du recensement; diagnostic; estimation empirique de Bayes; variances estimées; asymétrie; Lemme de Stein; échantillonnage.

1. INTRODUCTION

En ce qui concerne l'échantillonnage, le besoin d'élaborer des méthodes d'estimation exactes pour les petites régions représente un défi statistique de taille. Les estimations d'enquête directes comportent une variance trop grande pour être fiables relativement aux petites régions. Les techniques liées aux petites régions «améliorées» les estimations directes en les réduisant vers des valeurs lissées fondées sur un modèle. Des estimateurs simples de réduction ont été proposés par Purcell et Kish (1979). Dans un exposé innovant, Fay et Herriot (1979) ont montré qu'il peut en résulter un accroissement intéressant de la précision. Les articles de synthèse de Ghosh et Rao (1994) et de Singh, Gambino et Mantel (1994) témoignent nettement de

La mesure des erreurs des estimations régionales suscite un intérêt accru (voir Singh, Stukel et Pfeffermann 1998 et Booth et Hobert 1998). Le présent exposé propose l'estimation des erreurs quadratiques moyennes conditionnelles des estimateurs régionaux. L'erreur quadratique moyenne conditionnelle peut être estimée pour toutes les stratégies de réduction, y compris l'estimation empirique de Bayes et l'estimation théorique de décision (Purcell et Kish 1979). D'autres erreurs quadratiques moyennes, comme celle de Prasad et Rao (1990), et les propositions de Singh, Stukel et Pfeffermann (1998) axées sur les fréquences, mesurent la variabilité relativement au plan d'échantillonnage aussi bien que du modèle de lissage. L'erreur quadratique moyenne dont il est question dans le présent exposé est conditionnelle

en ce sens qu'elle mesure la variabilité relativement au plan d'échantillonnage pour une réalisation particulière du modèle de lissage. Cette caractéristique est intéressante puisque l'estimateur conditionnel reflète les conditions dans lesquelles l'enquête a été menée (voir Sæmstad, Swensson et Wretman 1992, ch. 7). L'inconvénient de cette propriété est est trop variable pour être utile concrètement. Lorsque le lissage est important, les estimateurs de l'erreur quadratique moyenne conditionnelle sont très instables. Il faut utiliser une évaluation inconditionnelle de la précision des estimateurs régionaux. Dans un tel cas, l'estimateur conditionnel proposé dans le présent exposé fournit néanmoins des renseignements utiles. On peut le considérer comme un outil diagnostique pour la comparaison des modèles de lissage. Il peut également servir d'élément pour la comparaison d'estimations de Monte Carlo des erreurs quadratiques moyennes inconditionnelles lorsqu'il n'existe pas de formules analytiques comme celle de Prasad et Rao (1990). Les auteurs ont voulu évaluer l'exactitude des estimateurs du sous-dénombrement provincial et infra-provincial du Recensement du Canada. Royce (1992) et Rivest (1995) ont examiné des substituts des estimations directes pour le sous-dénombrement provincial. Dick (1995) a appliqué des méthodes empiriques de Bayes aux estimations du sous-dénombrement infra-provincial. Ces deux exemples sont abordés à la section 5. Un estimateur de l'erreur quadratique moyenne conditionnelle est présenté à la section 2. Sa construction se fonde sur la version multivariée du lemme de Stein pour

¹ Louis-Paul Rivest et Eve Belmonte, Département de mathématiques et de statistiques, Université Laval, Sainte-Foy (Québec) Canada G1K 7P4.

- DUNCAN, G.J., et KALTON, G. (1987). Issues of design and analysis of surveys across time. *Revue Internationale de Statistique*, 55, 97-117.
- GOLDSSTEIN, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, 73, 43-56.
- GOLDSSTEIN, H. (1995). *Multilevel Statistical Models* (2ième édition). New York: Halstead.
- GOLDSSTEIN, H., HEALY, M.J.R., et RASBASH, J. (1994). Multilevel time series models with applications to repeated measures data. *Statistics in Medicine*, 13, 1643-1655.
- HARVEY, A.C. (1989). *Forecasting, structural time series models and the Kalman filter*. New York: Cambridge University Press.
- HERRIOT, R.A., et KASPRZYK, D. (1984). The survey of income and program participation. *Proceedings of the Social Statistics Section, American Statistical Association*, 107-116.
- JONES, R.H., et ACKERSON, L.M. (1990). Serial correlation in unequally spaced longitudinal data. *Biometrika*, 77, 721-731.
- JONES, R.H., et BOADIL-BOATING, F. (1991). Unequally spaced longitudinal data with AR(1) serial correlation. *Biometrics*, 47, 161-175.
- JONES, R.H., et VECCHIA, A.V. (1993). Fitting continuous ARMA models to unequally spaced spatial data. *Journal of the American Statistical Association*, 88, 947-954.
- JONES, R.H. (1993). *Longitudinal Data with Serial Correlation. A State-space Approach*. New York: Chapman and Hall.
- LAWLESS, J.F. (1999). Event history analysis and longitudinal surveys. Article présentée à la Conférence on Analysis of Survey Data, Southampton, United Kingdom.
- NATHAN, G. (1999). A Review of sample attrition and representativeness in three longitudinal surveys. GSS Methodology Series No. 13. London Office of National Statistics (ONS), United Kingdom.
- PFEEFFERMANN, D. (1993). The role of sampling weights when modeling survey data. *Revue Internationale de Statistique*, 61, 317-337.
- DUNCAN, G.J., et KALTON, G. (1987). Issues of design and analysis of surveys across time. *Revue Internationale de Statistique*, 55, 97-117.
- GOLDSSTEIN, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, 73, 43-56.
- GOLDSSTEIN, H. (1995). *Multilevel Statistical Models* (2ième édition). New York: Halstead.
- GOLDSSTEIN, H., HEALY, M.J.R., et RASBASH, J. (1994). Multilevel time series models with applications to repeated measures data. *Statistics in Medicine*, 13, 1643-1655.
- HARVEY, A.C. (1989). *Forecasting, structural time series models and the Kalman filter*. New York: Cambridge University Press.
- HERRIOT, R.A., et KASPRZYK, D. (1984). The survey of income and program participation. *Proceedings of the Social Statistics Section, American Statistical Association*, 107-116.
- JONES, R.H., et ACKERSON, L.M. (1990). Serial correlation in unequally spaced longitudinal data. *Biometrika*, 77, 721-731.
- JONES, R.H., et BOADIL-BOATING, F. (1991). Unequally spaced longitudinal data with AR(1) serial correlation. *Biometrics*, 47, 161-175.
- JONES, R.H., et VECCHIA, A.V. (1993). Fitting continuous ARMA models to unequally spaced spatial data. *Journal of the American Statistical Association*, 88, 947-954.
- JONES, R.H. (1993). *Longitudinal Data with Serial Correlation. A State-space Approach*. New York: Chapman and Hall.
- LAWLESS, J.F. (1999). Event history analysis and longitudinal surveys. Article présentée à la Conférence on Analysis of Survey Data, Southampton, United Kingdom.
- NATHAN, G. (1999). A Review of sample attrition and representativeness in three longitudinal surveys. GSS Methodology Series No. 13. London Office of National Statistics (ONS), United Kingdom.
- PFEEFFERMANN, D. (1993). The role of sampling weights when modeling survey data. *Revue Internationale de Statistique*, 61, 317-337.
- ZIMMERMAN, D.L., et NUNEZ-ANTON, V. (1997). Structured antedependence models for longitudinal data. In: *Modelling longitudinal and spatially correlated data: methods, applications and future directions*, (eds. T.G. Gregoire et coll.). Lecture Notes in Statistics, 22. New York: Springer Verlag, 62-76.
- WEBBER, M. (1994). The survey of labor and income dynamics: lessons learned in testing. *Proceedings of the Annual Research Conference, US Bureau of the Census*, 85-99.
- ZIMMERMAN, D.L., et NUNEZ-ANTON, V. (1997). Structured antedependence models for longitudinal data. In: *Modelling longitudinal and spatially correlated data: methods, applications and future directions*, (eds. T.G. Gregoire et coll.). Lecture Notes in Statistics, 22. New York: Springer Verlag, 62-76.
- RAO, J.N.K. (1999). Quelques progrès récents concernant l'estimation régionale fondée sur un modèle. *Techniques d'enquête*, 25, 199-212.
- SALTAS, W.H., et HARVILLE, D. A. (1981). Best linear recursive estimation for mixed linear models. *Journal of the American Statistical Association*, 76, 860-869.
- SKINNER, C.J., HOLT, D., et SMITH, T.M.F. (eds.) (1989). *Analysis of complex surveys*. Chichester: Wiley.
- SKINNER, C.J., et HOLMES, D. (1999). Random effects models for longitudinal survey data. Article présentée à la Conférence on Analysis of Survey Data, Southampton, United Kingdom.
- SURVEY RESEARCH CENTER (1984). User Guide to the Panel Study of Income Dynamics. Ann Arbor, Michigan: Inter-university Consortium for Political and Social Research.
- WEBBER, M. (1994). The survey of labor and income dynamics: lessons learned in testing. *Proceedings of the Annual Research Conference, US Bureau of the Census*, 85-99.
- ZIMMERMAN, D.L., et NUNEZ-ANTON, V. (1997). Structured antedependence models for longitudinal data. In: *Modelling longitudinal and spatially correlated data: methods, applications and future directions*, (eds. T.G. Gregoire et coll.). Lecture Notes in Statistics, 22. New York: Springer Verlag, 62-76.

trimestre qui précède) est un autre résultat intéressant qui se dégage du tableau 6. En outre, pour $q = 6$, la RMCI diminue à mesure que le nombre d'enregistrements par ménage augmente, comme l'explication l'utilisation de données observées pour d'autres membres du ménage. Enfin, dans les conditions du modèle, la RMCI est beaucoup plus faible pour les ménages à trois enregistrements que pour ceux à un ou à deux enregistrements, mais nous mentionnerons de nouveau qu'il n'existe trois enregistrements que pour deux ménages seulement. Le résultat inattendu tiré du tableau 6 est que, pour les ménages pour lesquels il n'existe qu'un seul enregistrement, la RMCI est un peu plus grande que pour $q = 6$ que pour $q = 5$ (on notera la corrélation assez forte et inexpliquée de 0,59 entre les valeurs corrigées calculées pour ces ménages à un intervalle de trois trimestres) et que pour $q = 2$ et $q = 5$, la RMCI pour les ménages pour lesquels on possède deux enregistrements est plus grande que la valeur correspondante pour les ménages pour lesquels on ne possède qu'un seul enregistrement. Ces anomalies ne sont pas inhabituelles dans le cas de données empiriques de taille assez petite et se manifestent encore plus fortement dans le cas du prédicteur naïf. (Le fait que, pour un nombre donné d'enregistrements par ménage, la RMCI calculée en appliquant le modèle pour $q = 5$ soit de même grandeur que les autres RQMI est rassurant, étant donné que, dans ce cas, les prédictions sont trois trimestres en avance.)

7. CONCLUSIONS ET EXTENSION DU MODÈLE

Les résultats présentés ici montrent qu'il est possible d'ajuster des modèles de série chronologique à des séries longitudinales très courtes et pour lesquelles des observations manquent. Le modèle utilisé ici est une extension du modèle linéaire type à deux niveaux en vertu duquel les effets aléatoires de premier et de deuxième niveaux évoluent de façon stochastique au fil du temps. Ce genre de modèle convient pour la modélisation des mesures longitudinales faites sur des populations à structure hiérarchique. Nous montrons que l'application de l'algorithme MCGIPP conjuguée à la pondération probabiliste type de la fonction de vraisemblance des séries chronologiques protège contre les effets de l'échantillonnage informatif.

Les modèles multinationaux sont souvent ajustés à des données discrètes, auquel cas ils contiennent des composantes non linéaires. En principe, la méthode d'estimation en deux étapes proposée ici peut être appliquée dans ces conditions également, même si, dans le cas de séries longitudinales très courtes, la gamme de modèles que l'on peut ajuster est manifestement limitée. De surcroît, une méthode commune d'estimations des paramètres inconnus du modèle dans le cas des données discrètes consiste à linéariser les composantes non linéaires à chaque itération de l'algorithme MGCI autour des estimations obtenues lors

BIBLIOGRAPHIE

de l'itération précédente, puis d'appliquer l'algorithme MCGI type pour calculer les estimations corrigées. Consulter Goldstein (1995) pour plus de précisions. Donc, il semble possible d'étendre sans grande difficulté l'algorithme MCGIPP au cas des données discrètes.

Dans le présent article, nous avons considéré l'estimation de la variance des estimateurs simples de la variance pour les estimateurs MCGIPP. Cependant, l'estimation de la variance des estimateurs obtenue par maximisation de la vraisemblance des séries chronologiques est plus problématique pour deux raisons. En premier lieu, la longueur éventuellement faible des séries longitudinales pourrait ne plus justifier l'utilisation de la matrice d'information ou ne plus permettre d'estimation stable, même si le nombre d'unités de deuxième niveau est grand. En deuxième lieu, les estimateurs MMN sont maintenus fixes lors de la maximisation de la vraisemblance, ce qui sous-entend que l'EMV est extrait des erreurs d'échantillonnage lors de l'estimation des paramètres MMN. Un moyen de résoudre ce problème consisterait à utiliser des méthodes de rééchantillonnage qui permettent de tenir compte de toutes les sources de variations dans le processus d'estimation.

Enfin, nous mentionnons une application importante du modèle proposé pour l'imputation de données manquantes. Dans un article qui paraîtra bientôt, Pfeffermann et Nathan (à paraître) illustrent la réduction importante de la variance d'imputation que permet de réaliser le modèle comparativement à des méthodes d'imputation plus classiques qui ne tiennent pas compte des effets communs des ménages.

- BELCHER, J., HAMPTON, J.S., et TUNNICLIFFE WILSON, G. (1994). Parametricization of continuous time autoregressive models for irregularly sampled time series data. *Journal of the Royal Statistical Society B*, 56, 141-155.
- BINDER, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-292.
- BINDER, D.A. (1998). Les enquêtes longitudinales: Pourquoi ces enquêtes diffèrent-elles de toutes les autres enquêtes? *Techniques d'enquête*, 24, 107-115.
- BRYANT, J., et DAY, R. (1991). Empirical Bayes analysis for systems of mixed models with linked autocorrelated random effects. *Journal of the American Statistical Association*, 86, 1007-1012.
- CHI, E.M., et REINSEL, G.C. (1989). Models for longitudinal data with random effects and AR(1) errors. *Journal of the American Statistical Association*, 84, 452-459.
- DIGGLE, P.J., LIANG, K.Y., et ZEGGER, S.L. (1994). *Analysis of Longitudinal Data*. Oxford: Clarendon Press.

données sur les ménages observés durant les trimestres antérieurs, durant lesquels la personne ne faisait pas partie de l'échantillon. On notera que, si l'on soustrait les effets fixes des observations originales, la distribution des valeurs corrigées ne dépend plus des trimestres de l'année civile. L'imputation pour le trimestre q est l'erreur de prévision correspondante qui, selon (3.1), est calculée comme étant

$$d_{hjq} = (r_{hjq} - z_{hjq}^{bq} - z_{hjq}^{bq-m} - e_{hjq}^{bq-m}) = r_{hjq}^{bq} - (z_{hjq}^{bq} + 1)' \alpha_{hjq}^{bq-m}$$

où $q = 2, 5, 6$ ou 7 est le prédicteur du vecteur d'état $\alpha_{hjq}^{bq} = (n_{hjq}, e_{hjq}^{bq})'$ si l'on se sert des données observées jusqu'au trimestre $q-m$, où $m = q - 1$ pour $q = 2, 6$ et $m = 3$ pour $q = 5$. On obtient le prédicteur $\hat{\alpha}_{hjq}^{bq-m}$ par application du filtre de Kalman avec les paramètres estimés correspondants (voir la section 2.3 et les équations 3.5 et 4.3).

Le tableau 6 montre la racine de la moyenne du carré des innovations (RMCI) selon le trimestre et le nombre d'enregistrements par ménage, tel que calculé conformément aux deux méthodes d'estimation (en se servant des valeurs paramétriques présentées au tableau 5). Aux fins de comparaison, nous présentons aussi la RMCI des innovations obtenues par prévision de la valeur corrigée pour le trimestre q au moyen de la valeur corrigée obtenue au trimestre précédent. On peut interpréter le prédicteur «naïf» $\hat{r}_{hjq}^{bq} = r_{hjq}^{bq-m}$ comme étant le prédicteur optimal dans les conditions du modèle de marche aléatoire simple $r_{hjq}^{bq} = r_{hjq}^{bq-m} + \text{erreur}$. Les moyennes des innovations pour les ménages possédant un enregistrement, $q = (2, 5, 6)$ sont $(0,68, 0,24, 0,301)$ pour les ménages possédant un enregistrement, $(1,24, -1,20, 0,60)$ pour les ménages possédant deux enregistrements et $(4,02, -5,82, 7,68)$ pour lesquels on possède trois enregistrements, mais on se souviendra que ces dernières valeurs ne se fondent que sur la moyenne de deux ménages. Les moyennes correspondantes des innovations empiriques calculées dans les conditions du modèle sont plus faibles en valeur absolue dans tous les cas.

Tableau 6

Racine de la moyenne du carré des innovations, selon le nombre d'enregistrements par ménage et le trimestre, dans le cas de deux méthodes d'estimations et du prédicteur naïf. Données de l'EPA

Nbre d'enr.	1	2	3
Trimestre	2	5	6
Méthode 1	11,54	11,16	12,26
Méthode 2	11,71	11,49	12,10
Préd. naïf	14,00	11,92	13,60

Le comportement des données analysées à la présente section est beaucoup plus imprévisible que celui des données utilisées pour l'étude en simulation produite conformément au modèle et nous ne pouvons prétendre que le modèle employé produit le meilleur ajustement possible (voir aussi plus bas). Néanmoins, les valeurs présentées au tableau 6 illustrent certaines caractéristiques importantes du

modèle. En premier lieu, mentionnons les résultats dans l'ensemble nettement meilleurs pour les prédicteurs du modèle que pour le prédicteur naïf $\hat{r}_{hjq}^{bq} = r_{hjq}^{bq-m}$, les deux méthodes d'estimations produisant une RMCI comparable. La supériorité du modèle tient au fait que, même si les autocorrélations de premier ordre des deux effets aléatoires des ménages utilisées pour les prédictions du modèle sont fortes en valeur absolue (très fortes pour la première composante), les autocorrélations des valeurs corrigées (les erreurs «totales») sont de taille modeste uniquement. Les autocorrélations de premier ordre des composantes aléatoires sont les coefficients d'autorégression correspondants (voir le tableau 5). Les autocorrélations des valeurs corrigées, $\text{Corr}(r_{hjq}^{bq}, r_{hjq}^{bq-m})$, $q = 2, 5, 6$; $m = 1$ pour $q = 2, 6$; $m = 3$ pour $q = 5$ sont, respectivement, $(0,46, 0,59, 0,51)$ pour les ménages à un seul enregistrement, $(0,48, 0,36, 0,45)$ pour les ménages à deux enregistrements et $(0,92, 0,43, 0,63)$ pour les ménages à trois enregistrements (d'après six enregistrements individuels).

Comme on l'a déjà mentionné, le fait qu'il n'existe qu'un seul enregistrement individuel pour la plupart des ménages pose des problèmes d'identifiabilité, puisque, pour ces ménages, il est impossible de faire la distinction entre les trois effets aléatoires. Le calcul des corrélations $\text{Corr}(r_{hjq}^{bq}, r_{hjq}^{bq-m})$, aux termes du modèle basé sur les estimations des paramètres du tableau 5, montre une bonne correspondance avec les corrélations calculées pour les ménages pour lesquels on possède deux enregistrements. Ce résultat montre, à son tour, que les estimateurs du tableau 5 sont dominés par ces observations et nous concluons que le modèle s'adapte le mieux aux observations obtenues pour les ménages pour lesquels on possède deux enregistrements. Cependant, on notera que les RMCI obtenues pour les autres tailles de ménages ne sont pas supérieures à celles calculées pour les ménages à deux enregistrements (voir aussi plus bas). Il est important de mentionner à cet égard que si l'on avait agrégé les données recueillies pour toutes les personnes observées durant un trimestre particulier de l'année civile, il aurait été impossible de tenir compte des effets aléatoires des ménages, ce qui aurait abouti à de moins bonnes prévisions des observations individuelles. À cet égard, voir la discussion dans l'introduction. (La modélisation des données agrégées est assez compliquée dans ce cas, puisque pour chaque trimestre de l'année civile, l'échantillon comprend quatre panels distincts, tels que définis par le nombre de fois qu'une personne fait partie de l'échantillon. Autrement dit, les modèles valables pour ces panels varient selon le nombre d'observations disponibles pour chaque panel.)

Le fait que les RMCI obtenues en appliquant le modèle sont généralement plus faibles pour $q = 6$ que pour $q = 2$, comme l'explique l'utilisation d'un plus grand nombre de données observées pour la même personne dans le processus prévisionnel (un plus grand nombre de données observées pour estimer les effets aléatoires durant le

la méthode 2 sont les mêmes que les estimateurs correspondants du tableau 3.

Tableau 4

Moyennes, écarts-types (ET) et statistiques <i>t</i> des estimateurs dans le cas de deux méthodes d'estimation. Échantillonnage informatif, MMIN pondéré, fonction de vraisemblance non pondérée									
Méthode 1					Méthode 2				
paramètre	valeur réelle	ET	statistique	moyenne ET	paramètre	valeur réelle	ET	statistique	moyenne ET
A_{11}	0,500	0,468	0,09	-3,477	0,453	0,10	-4,569	3,197	0,11
A_{22}	0,700	0,742	0,11	3,948	0,737	0,11	3,197	1,571	0,19
Δ_{11}^2	1,067	1,060	0,11	-0,598	1,040	0,17	-1,560	1,571	0,19
Δ_{22}^2	0,980	1,008	0,11	2,449	1,010	0,19	1,571	0,894	0,02
ρ	0,400	0,407	0,02	3,021	0,402	0,02	0,894	0,298	0,01
σ_e^2	0,298	0,298	0,01	1,013	0,294	0,01	-3,340		

Le résultat intéressant qui se dégage du tableau 4 est que

les estimateurs de A_{11} et A_{22} sont maintenant entachés d'un biais non négligeable, contrairement aux estimateurs correspondants du tableau 3. Ce résultat s'explique comme suit. Dans le cas du plan d'échantillonnage informatif, l'espérance des effets aléatoires $u_{h,1}$ correspondants aux ménages h compris dans l'échantillon est inférieure à 0, $E(u_{h,1} | h \in s) < 0$, et, par conséquent, l'initialisation du filtre de Kalman par l'espérance de population ($E u_{h,1} = 0$, équation 4.4) produit des estimateurs biaisés. Par ailleurs, si l'on pondère les contributions à la vraisemblance L_h par l'inverse de la probabilité de sélection dans l'échantillon, les proportions de ces contributions L_h correspondant aux effets aléatoires dont la valeur est supérieure ou inférieure à celle prévue par le modèle sont compensées par les proportions de population et, donc, l'utilisation de la valeur prévue par le modèle pour l'initialisation ne biaise pas le processus d'estimation. Comme on l'a remarqué pour les tableaux précédents, la valeur de l'ET des estimateurs non biaisés du tableau 4 est beaucoup plus faible que celle de l'ET des estimateurs pondérés correspondants du tableau 3.

6. APPLICATION DU MODÈLE AUX DONNÉES DE L'EPA

Nous avons ajusté le modèle défini par (3.1) et (3.4) à un ensemble de données empiriques extrait des données recueillies dans le cadre de l'EPA israélienne pour Jérusalem de 1990 à 1994. Les données contiennent les enregistrés complets pour 567 personnes dans 475 ménages, chaque personne étant observée au cours de quatre trimestres, conformément au plan de renouvellement décrit plus haut et utilisé pour l'étude de renouvellement. Parmi les 475 ménages, 385 possèdent un enregistrement individuel, 88 possèdent deux enregistrements individuels et 2 seulement possèdent trois enregistrements individuels. La variable observée est y = nombre d'heures de travail durant la semaine qui a précédé l'entrevue, ($\bar{y} = 39,8$, s.d(y) = 14,8;

calculé sur toutes les personnes et sur tous les trimestres). Les variables auxiliaires de niveau individuel sont x_1 = nombre d'années d'études, ($\bar{x}_1 = 13,4$, s.d(x_1) = 4,8) et x_2 = sexe, (41 % de femmes). Les variables auxiliaires au niveau des ménages sont $z_1 = 1$ et z_2 = nombre de personnes occupées dans le ménage ($\bar{z}_2 = 1,48$, s.d(z_2) = 0,56). Nous avons estimé les paramètres du modèle par les deux méthodes décrites à la section 4. Les poids d'échantillonnage appliqués à ces données sont fort semblables d'un ménage à l'autre et d'une personne à l'autre, de sorte que nous avons calculé uniquement les estimateurs non pondérés. L'algorithme MCGI produit des estimations négatives de la variance pour certains trimestres et l'on a donné à ces estimations une valeur nulle lors du calcul de la moyenne des estimations de la variance produite par la méthode 2. On n'a pas calculé la moyenne des estimations trimestrielles des coefficients fixes du modèle, puisqu'ils varient considérablement au cours de la période de cinq ans. Les estimations de la variance et des coefficients d'autorégression calculées par les deux méthodes sont présentées au tableau 5 en se servant de la même notation que dans les tableaux précédents. Les deux ensembles d'estimations ne diffèrent pas fortement, sauf dans le cas de l'estimateur de Δ_{22}^2 qui, comme on l'a déjà mentionné, a une valeur négative pour certaines exécutions de l'algorithme MCGI. On notera à cet égard que, pour la plupart des ménages, on ne possède qu'un seul enregistrement individuel (voir plus haut) et que, pour presque tous ces ménages, $z_2 = 1$. Cette situation complique le processus d'estimation, puisque, pour ces ménages, il est impossible de faire la distinction entre l'effet de premier niveau (individu) et les deux effets des ménages, qui sont pareillement confondus. (On remarquera que la somme des deux dernières variances est la même dans le cas des deux méthodes). Comme on le note plus bas, les estimateurs présentés au tableau 5 sont dominés par les observations obtenues pour les ménages pour lesquels on dispose de deux enregistrements individuels.

Tableau 5
Estimations de la variance et des coefficients d'autorégression dans le cas de deux méthodes d'estimation.
Données de l'EPA

Paramètre	A_{11}	A_{22}	Δ_{11}^2	Δ_{22}^2	ρ	σ_e^2
Méthode 1	0,915	-0,606	73,88	2,541	0,242	102,306
Méthode 2	0,976	-0,548	56,88	14,753	0,448	101,001

Dans le cas du plan d'échantillonnage de l'EPA israélienne, chaque enregistrement individuel comprend quatre observations recueillies lors des trimestres 1, 2, 5 et 6, le trimestre 1 étant défini comme le premier trimestre de l'année civile t durant lequel la personne est incluse dans l'échantillon. Afin d'évaluer le pouvoir prédictif du modèle, nous calculons, pour chaque enregistrement individuel (h, j) les innovations empiriques lorsque l'on prédit les valeurs corrigées $r^{hjg} = (y^{hjg} - x^{hjg}_1 \hat{\gamma} - z^{hjg}_2 \hat{\beta})$ d'après les

Il est important de mentionner que l'ET des estimateurs pondérés présentes au tableau 3 est toujours supérieure à l'ET correspondante des estimateurs non pondérés présentés au tableau 2. Comme l'a souligné l'un des examinateurs, ce résultat sous-entend que les racines des erreurs quadratiques moyennes (REQM) empiriques des estimateurs non pondérés du tableau 2 sont en fait plus grandes que les REQM empiriques des estimateurs correspondants du tableau 3. Cependant, ce résultat tient aux tailles assez faibles d'échantillon utilisées pour la présente étude. Dans le cas de plus grands échantillons (grand nombre de ménages et de personnes dans les ménages), l'REQM est dominée par le biais qui, contrairement à la variance, ne diminue pas à mesure que l'augmentation la taille de l'échantillon. Par conséquent, il est évident que, à mesure qu'augmente la taille de l'échantillon, l'REQM des estimateurs pondérés devient plus petite que l'REQM des estimateurs non pondérés. Le fait que la variance des estimateurs à pondération probabiliste soit plus grande que celle des estimateurs correspondants non pondérés est un phénomène bien connu observé lors de nombreuses autres études. Pour une discussion approfondie et une bibliographie, consulter Pfeffermann (1993).

Tableau 3

Moyennes, écarts-types (ET) et statistique t des estimateurs dans le cas de deux modèles d'estimations. Echantillonnage informatif, estimateurs pondérés			
Méthode 1		Méthode 2	
paramètre	valeur réelle	paramètre	valeur réelle
γ_1	6,000	5,997	0,04
γ_2	-2,000	-2,000	0,05
γ_3	1,000	0,978	0,14
ν_1	2,000	2,019	0,14
ν_2	0,500	0,490	0,15
A_{11}	0,700	0,699	0,17
A_{22}	1,067	1,055	0,17
Δ_{11}	0,980	1,023	0,19
Δ_{22}	0,400	0,401	0,04
σ_e^2	0,298	0,297	0,01

Echantillonnage informatif, estimateurs pondérés

Donc, nous constatons que le biais relatif absolu qui entache l'estimation de ν_1 est d'environ 27% et nous observons aussi un biais relatif important pour les estimateurs A_{11} et Δ_{11} . (Le modèle défini par (3.1) peut être réécrit sous la forme $y_{hi} = x_{hi}^{\nu_1} \nu_1 + z_{hi}^{\nu_2} \nu_2 + e_{hi}$ où $u_{hi} = u_{hi}^{\nu_1} + \nu_1$, de sorte que, pour $\nu_1 = \nu$, comme dans le cas du modèle de simulation, $\nu_1 = E(u_{hi}^{\nu_1})$). On notera que les trois biais sont négatifs, situation due au fait que le mécanisme de sélection utilisé pour la présente étude comprend le suréchantillonnage des personnes pour lesquelles les observations contiennent des effets aléatoires négatifs $u_{hi}^{\nu_1}$. Dans ce cas de nouveau, les deux méthodes d'estimation donnent des résultats fort semblables.

Tableau 2

Moyennes, écarts-types (ET) et statistique t des estimateurs dans le cas de deux modèles d'estimation. Echantillonnage informatif, estimateurs non pondérés									
Méthode 1					Méthode 2				
paramètre		valeur réelle		t	paramètre		valeur réelle		t
γ_1	6,000	5,998	0,02	-0,768	γ_1	6,000	5,998	0,02	-0,768
γ_2	-2,000	-2,000	0,03	0,104	γ_2	-2,000	-2,000	0,03	0,104
ν_1	1,000	0,728	0,09	-34,385	ν_1	1,000	0,728	0,09	-34,385
ν_2	2,000	2,005	0,09	0,564	ν_2	2,000	2,005	0,09	0,564
A_{11}	0,500	0,438	0,09	-6,742	A_{11}	0,500	0,434	0,09	-7,453
A_{22}	0,700	0,738	0,09	4,078	A_{22}	0,700	0,735	0,09	3,941
Δ_{11}	1,067	0,995	0,09	-7,766	Δ_{11}	1,067	0,994	0,09	-7,883
Δ_{22}	0,980	1,003	0,10	2,352	Δ_{22}	0,980	0,987	0,10	0,698
p	0,400	0,407	0,02	3,184	p	0,400	0,405	0,02	2,218
σ_e^2	0,298	0,298	0,01	0,644	σ_e^2	0,298	0,296	0,01	-1,800

Le tableau 3 montre les résultats obtenus par application de l'algorithme MCGIPP pour l'estimation des paramètres du MNM (section 2.2) et par pondération des fonctions individuelles de vraisemblance de la série chronologique $\log(L_h) = -\{1/2 T_h n_h \log(2\pi) + 1/2 \sum_{i=1}^{T_h} \log |F_{hi}^m| + 1/2 (X_{hi}^m - \bar{X}_{hi}^{m-1})' F_{hi}^{m-1} (X_{hi}^m - \bar{X}_{hi}^{m-1})\}$ par les poids d'échantillonnage des ménages $w_h = 1 / \text{Pr}(h \in s)$, en se servant des mêmes 100 échantillons que ceux observés pour produire le tableau 2. La pondération des contributions à la vraisemblance par l'inverse de la probabilité d'inclusion dans l'échantillon est une application de la méthode de la pseudo-vraisemblance souvent recommandée pour ajuster les modèles à un seul niveau à des données transversales; à cet égard, consulter, par exemple, Binder (1983), Skinner et coll. (1989) et Pfeffermann (1993). Comme le montre ce tableau, l'application de l'algorithme MCGIPP et la pondération de la fonction de vraisemblance éliminent les biais importants observés au tableau 2, malgré l'initialisation incorrecte du filtre de Kalman dans le cas de séries très courtes. (Voir la discussion au Commentaire 1 à la fin de la section 4.) De nouveau ici, les deux méthodes d'estimation donnent des résultats fort semblables, produisant dans chaque cas un estimateur entaché d'un biais relatif très faible.

Comme on l'a exposé au Commentaire 1 à la section 4, l'échantillonnage informatif fausse la distribution transversale des observations d'échantillon et l'initialisation du filtre de Kalman, mais n'a aucun effet sur les distributions conditionnelles des effets aléatoires de premier et de deuxième niveaux définis par (3.4). Donc, il est intéressant de déterminer si l'utilisation de l'algorithme MCGIPP pour estimer les paramètres du modèle transversal sans pondérer de la même façon la fonction de vraisemblance de la série chronologique permet de neutraliser le biais. Le tableau 4 montre les résultats obtenus dans ce cas sur les mêmes échantillons que ceux utilisés pour produire les données des tableaux 2 et 3. Les estimateurs des coefficients vectoriels fixes $\beta' = (\gamma', \nu', v', \nu', \nu')'$ sont les mêmes qu'au tableau 3 et, par conséquent, ne sont pas présentés de nouveau. On notera que les estimateurs de Δ_{11} , Δ_{22} et σ_e^2 obtenus en appliquant

personnes dans les ménages sélectionnés.

C2) Échantillonnage informatif

Les valeurs de population ont été produites pour des panels de 55 ménages. On a sélectionné les ménages avec effet aléatoire $n_{h,1} > 0$ (la valeur du premier effet aléatoire au premier point dans le temps) avec une probabilité égale à 1 et les ménages avec effet aléatoire $n_{h,1} < 0$ indépendamment (échantillonnage de Poisson) avec une probabilité égale à 0,1. Toutes les personnes faisant partie d'un ménage échantillonné ont été observées. Ce plan d'échantillonnage produit une taille prévue d'échantillon de 30 ménages par panel et des tailles prévues d'échantillon de $n = 120$ ménages et $n = 360$ personnes par trimestre, tailles équivalentes à celles obtenues dans le cas du plan d'échantillonnage C1.

plan d'échantillonnage C1.

Commentaire 2: Il convient de souligner que, même si l'on

considère les deux séries de paramètres du modèle. Les paramètres des transitions de ménage pour l'estimation de panel sont observés durant un trimestre uniquement. En tout, ceci représente l'observation de 13 transitions de panel et environ 390 transitions de ménage pour l'estimation des paramètres de la série chronologique. (Par transition de panel, nous entendons que le même panel est observé lors de deux cycles. Pour trois de ces transitions de panel, il existe un décalage de deux trimestres entre les deux observations.) Nous nous reportons à cette structure d'échantillon pour évaluer l'estimation des paramètres de

modèle de série chronologique.

Pour chacun des plans d'échantillonnage C1 et C2, nous avons répété 100 fois le processus entier de production de valeurs de population et de sélection de l'échantillon, en sélectionnant un échantillon pour chaque population. Puis, nous avons appliqué à chaque échantillon les deux méthodes d'estimation décrites à la section 4. Les simulations ont été exécutées au moyen du logiciel Gauss. Pour procéder à la maximisation de la vraisemblance, nous avons appliqué la méthode d'optimisation numérique OPTIMUM.

5.2 Résultats

Les résultats de l'étude en simulation sont résumés aux tableaux 1 à 4 sous forme de moyennes sur les 100 échantillons sélectionnés conformément aux deux plans d'échantillonnage. Chaque tableau contient les estimations moyennes des paramètres du modèle, les écarts-types (ET) empiriques des estimateurs et la statistique t classique obtenue en divisant la différence entre l'estimation de la moyenne et la valeur réelle du paramètre par l'écart-type, calculée comme

Le résultat le plus important qui se dégage du tableau 1 est peut-être que, dans des conditions d'échantillonnage non informatif, il est effectivement possible de bien ajuster des modèles simples, mais non triviaux, de séries chronologiques à des séries longitudinales très courtes, à condition que le nombre de séries observées soit suffisamment grand. Le modèle n'est pas trivial, car, même si l'on soustrait les effets fixes, la variable réponse, ou variable dépendante, est égale à la somme des trois processus AR (1). Le fait que huit des 11 panels aient été observés au moins deux fois, ce qui donne en tout 13 transitions de panel, dont trois avec un décalage de deux trimestres, renforce encore cette conclusion. Voir le Commentaire 2 à la fin de la section 5.

Tableau 1
Moyennes, écart-types (ET) et statistiques t des
estimateurs pour deux modèles d'estimation.
Echantillonage non informatif

Méthode 1		Méthode 2	
Paramètre	vaaleur réelle	statistique	statistiques
γ_1	6,000	6,002	0,677
γ_2	-2,000	-2,000	0,03
ν_1	1,000	0,989	-1,357
ν_2	2,000	2,008	0,08
A_{11}	0,500	0,497	-0,391
A_{22}	0,700	0,696	-0,532
Δ_{11}^2	1,067	1,054	-1,668
Δ_{22}^2	0,980	0,991	1,042
ρ	0,400	0,398	-0,937
σ_z^2	0,298	0,298	0,297
			0,01
			-0,062
			0,02
			-0,937
			0,397
			0,02
			-1,637
			0,01
			-1,382

L'évaluation de la performance des deux ensembles d'estimateurs présentée au tableau 1 montre, si l'on s'en tient à la statistique χ^2 que, dans le cas de la méthode 1, tous les estimateurs sont fortement non significatifs et que, dans le cas de la méthode 2, seul l'estimateur de Δ_1 est significatif. À noter que, même dans ce cas, le biais relatif absolu est d'environ 2% et, puisque les EMV des paramètres des séries chronologiques ne sont généralement pas strictement sans biais, un petit biais comme celui-ci dans l'un d'un des cas des deux méthodes, les erreurs-types des estimateurs moyens sont fort semblables, résultats que l'on observe aussi dans les autres tableaux.

Considérons maintenant le cas de l'échantillonnage informatif. Le tableau 2 montre les résultats obtenus si l'on ne tient pas compte du processus d'échantillonnage informatif, en appliquant les mêmes méthodes d'estimation que dans le cas de l'échantillonnage non informatif. Comme l'indique très clairement ce tableau, certaines estimations paramétriques sont fortement significatives, particulièrement les estimateurs des paramètres qui indexent le modèle de série chronologique des effets aléatoires μ_{it} , qui définissent les probabilités de sélection dans l'échantillon.

MCGIFF. Donc, pour chaque échantillon S_j , on se sert de l'algorithme MCGIPP plutôt que de l'algorithme MCGI (MMN).

Commentaire 1 : La sélection informative des unités de premier et de deuxième niveaux n influe pas sur la distribution conditionnelle des effets aléatoires définis par (3.4). Donc, la distribution de n_{h1} et de e_{h1} pourrait être fortement faussée par la sélection de l'échantillon au temps $t = 1$, mais cette situation n'a aucun effet sur les distributions de $n_{h2} | n_{h1}$, ni de $e_{h2} | e_{h1}$. Étant donné cette propriété, le calcul de la vraisemblance à la deuxième étape reste le même, mais il faut s'assurer de procéder à l'initialisation appropriée du filtre de Kalman. On initialise ce dernier, tel qu'il est défini par (4.4), au moyen des moyennes et des variances non conditionnelles des effets aléatoires dans les conditions du modèle, mais, au temps $t = 1$, les moments valables pour les unités de l'échantillon peuvent différer, à cause des effets d'échantillonnage. Il est bien connu que, pour des séries suffisamment longues et dans certaines conditions de régularité, les estimations calculées par maximisation de la fonction de vraisemblance ne sont pas sensibles à la méthode d'initialisation, mais que, dans le cas des séries courtes, l'initialisation incorrecte dans des conditions d'échantillonnage informatif pourrait fausser le processus d'estimation. Néanmoins, comme on l'illustre à la section 5, dans le cas d'un nombre modéré d'enregistrements longitudinaux, même très courts (au plus quatre observations dans notre application), la pondération des contributions à la vraisemblance par l'inverse de la probabilité d'inclusion dans l'échantillon (application de la méthode du pseudo-maximum de vraisemblance) produit des estimateurs non biaisés pour tous les paramètres du modèle de la série chronologique.

5. RÉSULTATS DE LA SIMULATION

Nous présentons ici les résultats d'une étude de Monte Carlo réalisée pour évaluer la performance de diverses méthodes d'estimation décrites à la section 4 dans le cas de plans d'échantillonnage avec renouvellement informatif et non informatif.

5.1 Description de l'étude en simulation

A) Production des données sur la population et plan de renouvellement de l'échantillon

On a produit les valeurs de population pour les personnes (unités de premier niveau) dans les ménages (unités de deuxième niveau) au moyen du modèle défini par (3.1) et (3.4) (voir plus bas). On a déterminé au hasard le nombre de personnes n_h observées dans un ménage h , les valeurs possibles étant 2, 3 ou 4. On a produit un nouveau panel de ménages pour chaque

B) Modèle de population

Le modèle utilisé pour produire les valeurs de y pour un ménage h particulier est défini par (3.1) et (3.4) où $x'_{h1} \equiv (x_{h1}, x_{h2})$ et $z'_{h1} \equiv (z_{h1}, z_{h2})$, de sorte que les valeurs des covariables sont fixes au cours du temps. Les valeurs de x sont produites indépendamment d'après la distribution uniforme $U[1, 2]$. Les valeurs z_{h2} ont été produites d'après la distribution uniforme $U[1, 5]$. Afin de simplifier la présentation et d'évaluer les résultats, nous avons également supposé que les coefficients du modèle étaient invariables, de sorte que $\gamma_i = \gamma = (6, -2)'$ et $v_i = v = (1, 2)'$. Les termes d'erreur aléatoire ont été générés indépendamment entre les ménages au moyen du modèle (3.4) où $A_2 = \text{diag}[0.5, 0.7]$, $\Delta = \text{diag}[0.8, 0.5]$, $\rho = 0.4$ et $\sigma^2_2 = 0.25$. On notera, compte tenu de (4.1), que $\text{Var}(u_{h1}) = \Delta' = \text{diag}[1.067, 0.980]$ et $\text{Var}(e_{h1}) = \sigma^2_2 = 0.298$.

C) Sélection de l'échantillon

Nous considérons deux plans distincts d'échantillonnage.

C1) Échantillonnage non informatif

Les valeurs de population ont été produites pour des panels de 30 ménages, où tous les ménages faisant partie d'un panel donné ont été sélectionnés dans l'échantillon et observés conformément au plan de renouvellement de l'échantillon décrit en A. Le nombre total de ménages échantillonnés à chacun des trimestres 6 à 11 est par conséquent égal à $m = 120$. Toutes les personnes faisant partie d'un ménage donné ont été observées, ce qui donne une taille prévue d'échantillon de $n = 360$ personnes à chaque trimestre. Ce plan d'échantillonnage correspond à

La méthode commence par l'ajustement du MMN défini par (3.1) à chaque échantillon S_i séparément, pour obtenir les estimations MCGI des effets fixes chronologiques $\beta_i = [\gamma_i', \nu_i']'$ et la variance des effets aléatoires u_{hi} et e_{hi} . On notera que, étant donné (3.4),

$$\text{Var}(u_{hi}) = \Delta^* = (I - A_2)^{-1} \Delta;$$

$$\text{Var}(e_{hi}) = \sigma_e^2 = (1 - p_2) \sigma_e^2 \quad (4.1)$$

Afin d'appliquer le filtre de Kalman et de calculer la vraisemblance, il faut fixer les valeurs initiales de α_{i10} et P_{i10} . L'exercice est simple dans le cas du modèle examiné ici, puisque $\alpha_{hi} = [u_{hi}', e_{hi}']'$ est stationnaire, avec une moyenne nulle et une matrice de covariance définie par (4.1). Donc, on commence le filtre en fixant

$$\alpha_{hi10} = E(u_{hi}', e_{hi}') = 0;$$

$$P_{hi10} = \text{Var}[u_{hi}', e_{hi}']$$

$$= \text{diag}\{(I - A_2)^{-1} \Delta, \sigma_e^2 (I - p_2)^{-1} I^{n_h}\}. \quad (4.4)$$

Si l'on utilise des relations bien connues qui tiennent pour les modèles AR(1), l'utilisation de cette étape produit les estimations $\{\beta_i', \Delta_i', \sigma_e^2\}$ pour $\{\beta_i', \Delta_i', \sigma_e^2\}$, respectivement. Dans les conditions du modèle, les variances réelles (Δ^*, σ_e^2) ne varient pas en fonction du temps et, si l'on suppose que l'effectif de l'échantillon à différents points dans le temps est assez constant, on peut calculer la moyenne des estimations Δ_i' et σ_e^2 pour obtenir les estimations uniques

$$\bar{\Delta}^* = \sum_{i=1}^I \Delta_i' / I; \quad \bar{\sigma}_e^2 = \sum_{i=1}^I \sigma_e^2 / I. \quad (4.2)$$

À la deuxième étape, on estime les autres paramètres en maximisant la vraisemblance du modèle combiné défini par (3.3), (3.5) et (3.6), en maintenant fixes, à leur valeur estimée, les paramètres calculés à la première étape. Puisque les observations faites sur des unités de deuxième niveau distinctes sont indépendantes, le logarithme du rapport de vraisemblance prend la forme $\log(L) = \sum_{h=1}^H \log(L_h)$ où L_h est la contribution de l'unité h de deuxième niveau à la vraisemblance, est défini par (2.6) où l'indice h est ajouté à toutes les composantes pour faire la distinction entre des unités de deuxième niveau différentes. Comme on l'a fait remarquer plus haut, le nombre de points dans le temps auxquels les unités de deuxième niveau sont observées et les périodes durant lesquelles les observations sont faites peuvent varier d'une unité à l'autre, si bien que, dans (2.6), on doit également modifier la notation T pour qu'elle devienne T_h .

Si l'on ajuste le modèle aux données obtenues conformément à un plan de sondage avec renouvellement du panel, comme dans l'étude empirique présentée ici, une modification supplémentaire est nécessaire pour tenir compte des périodes intermédiaires sans observation. Par exemple, dans le cas de l'EPA israélienne décrite dans l'introduction, dont le plan de renouvellement prévoit la présence d'une unité dans l'échantillon pendant deux trimestres, son exclusion de l'échantillon pendant deux trimestres, puis de nouveau sa présence dans l'échantillon pendant deux trimestres, $T_h = 4$, mais les équations de transition de $t = 2$ à $t = 3$ (le trimestre suivant d'observation) doivent être modifiées pour tenir compte des deux points dans le temps pour lesquelles des observations manquent. Des substitutions répétées dans (3.5) produisent les expressions qui suivent:

$$\bar{\Delta}^* = (1 - A_2)^{-1} \bar{\Delta}; \quad \bar{\sigma}_e^2 = (1 - p_2)^{-1} \bar{\sigma}_e^2. \quad (4.5)$$

Méthode 2: Les seuls paramètres estimés à la deuxième étape sont les coefficients de l'équation d'autorégression p, A_1, A_2 (équation 3.4). On notera que, par cette méthode, les variances Δ et σ_e^2 sont fixées dans l'expression de la vraisemblance comme étant $\bar{\Delta} = (I - A_2)^{-1} \bar{\Delta}^*$ et $\bar{\sigma}_e^2 = (1 - p_2) \bar{\sigma}_e^2$ en utilisant (4.1), où $\bar{\sigma}_e^2$ et $\bar{\Delta}^*$ sont définies par (4.2).

Les méthodes d'estimation décrites jusqu'ici reposent sur l'hypothèse implicite que l'échantillonnage n'est pas informatif. Comme on l'a exposé dans l'introduction, les enquêtes à plan de sondage complexe s'appuient souvent sur un échantillonnage avec probabilités inégales qui pourraient être corrélées aux valeurs de la variable réponse. Le cas échéant, le modèle valable pour les données d'échantillon peut différer de celui qui s'applique à la population. La méthode d'estimation en deux étapes proposées peut être adaptée de façon à éviter l'effet de l'échantillonnage informatif. Pour cela, on applique la méthode de pondération décrite à la section 2.2 à la première étape, en remplaçant l'algorithme itératif MCGI par l'algorithme

Par exemple, dans le cas de l'EPA trimestrielle israélienne décrite dans l'introduction, les personnes font partie de l'échantillon quatre trimestres en tout sur une période de six trimestres, situation qui limite manifestement la gamme de modèles de série chronologique que l'on peut supposer applicables aux effets aléatoires.

Les modèles AR(1) définis par (3.4) peuvent être représentés concisément par

$$\alpha_m = G_h \alpha_{h,t-1} + \eta_m, \quad h = 1, \dots, m, \quad (3.5)$$

où

$$G_h = \begin{bmatrix} A & 0 \\ 0 & pI_{n_h} \end{bmatrix}, \quad \eta_m = \begin{bmatrix} \varepsilon_m \\ \delta_m \end{bmatrix},$$

$$\eta_m \sim N(0, \bar{O}_h), \quad \bar{O}_h = \begin{bmatrix} \Delta & 0 \\ 0 & \sigma^2 I_{n_h} \end{bmatrix}. \quad (3.6)$$

En représentant le modèle proposé au moyen des équations (3.3), (3.5) et (3.6) et en posant que $Z_m^h = L_m^h$,

$H_m^h = 0$, il est facile de voir que ce modèle appartient à la catégorie des modèles d'espace d'états présentes à la section 2.3 sans erreur résiduelle dans l'équation de mesure. Le modèle est défini pour des unités h de deuxième niveau distinctes, mais, à l'encontre de l'analyse classique des séries chronologiques, où les données correspondent à une longue série unique, les données correspondent ici à un grand nombre de séries courtes (longitudinales) indépendantes observables durant diverses périodes. On notera que la matrice de transition, G_h , et la matrice de covariance, \bar{O}_h , dépendent de h par le biais de la taille de l'échantillon de deuxième niveau n_h , mais ne varient pas en fonction du temps. Dans les cas où les tailles de deuxième niveau varient avec le temps (par exemple, à cause de données manquantes), ces matrices varient en conséquence.

4. ESTIMATION DES PARAMÈTRES DU MODÈLE

En principe, on peut maximiser la fonction de vraisemblance valable pour le modèle défini par (3.3), (3.5) et (3.6) en vue d'obtenir les estimateurs du maximum de vraisemblance (EMV) de tous les paramètres inconnus du modèle. Cependant, le nombre de paramètres estimés est habituellement très grand, situation qui peut compliquer les calculs et produire des estimateurs statistiquement instables. Par exemple, même pour $p = q = 2$ et $T = 10$, il existe déjà 46 paramètres inconnus. Nous proposons donc une méthode d'estimation en deux étapes qui se fonde sur un modèle multiniveaux (MMN) pour les «paramètres transversaux» et sur un modèle d'espace d'états pour les «paramètres de série chronologique». Cette méthode permet en outre d'appliquer une pondération appropriée pour éviter les effets de l'échantillonnage informatif.

$$Y_m^h = X_m^h \gamma + Z_m^h \nu + I_{n_h} \varepsilon_m, \quad (3.2)$$

où $X_m^h = [x_{h1}^m, \dots, x_{hn_h}^m]'$, $X_m^h = [x_{h1}^m, \dots, x_{hn_h}^m]'$, $Z_m^h = [z_{h1}^m, \dots, z_{hn_h}^m]'$, où \otimes définit le produit de Kronecker. La représentation (3.2) de la matrice peut s'écrire concisément comme suit

$$Y_m^h = \tilde{X}_m^h \beta + \tilde{Z}_m^h \alpha_m, \quad (3.3)$$

où $\tilde{X}_m^h = [X_m^h, Z_m^h]$, $\tilde{Z}_m^h = [Z_m^h, I_{n_h}]$, $\beta = [\gamma', \nu']'$, $\alpha_m = [e_m^h, \varepsilon_m^h]'$.

Ensuite, nous modélisons la relation entre les coefficients vectoriels et les effets aléatoires de la série chronologique. Nous supposons que les vecteurs β , $t = 1, 2, \dots$ sont fixes sans préciser la façon dont ils évoluent au cours du temps. Cette hypothèse n'est habituellement pas restrictive, car, en général, dans les applications pratiques, la taille globale de l'échantillon à n'importe quel point dans le temps est suffisamment grande pour permettre une estimation exacte des coefficients vectoriels sans devoir recourir à des renseignements auxiliaires au cours du temps. En ce qui concerne les effets aléatoires de deuxième et de premier niveau, nous posons une relation autorégressive de premier ordre [AR(1)] de la forme

$$\eta_m^h = A \eta_{h,t-1} + \delta_m^h; \quad e_m^h = \rho e_{h,t-1} + \varepsilon_m^h \quad (3.4)$$

où A est une matrice $(q \times q)$ de coefficients fixes, ρ est une grandeur scalaire fixe et $\delta_m^h \sim N(0, \sigma_\delta^2 I_{n_h})$, $\varepsilon_m^h \sim N(0, \sigma_\varepsilon^2 I_{n_h})$ sont des bruits blancs gaussiens indépendants. Le modèle défini par (3.4) est assez simple et, pour le simplifier encore davantage, nous supposons que A et Δ sont diagonales, donc que les effets aléatoires de deuxième niveau sont indépendants. Nous supposons aussi que $|\rho| < 1$ et $|A_{kk}| < 1$ pour toutes les valeurs de k pour garantir la stationnarité. En principe, on pourrait considérer des modèles plus complexes, mais il faut insister sur le fait que, contrairement au cas de l'analyse classique (agrégée) des séries chronologiques, les observations longitudinales ne peuvent être recueillies que sur une très courte période, si bien que les modèles auxquels sont intégrées des valeurs retardées d'ordre élevé pourraient ne plus être applicables.

stochastiques, bien qu'elles puissent varier au cours du temps, comme cela est le cas pour les coefficients vectoriels β_i . On remarquera que l'on peut inclure ces derniers vecteurs dans les vecteurs d'état en représentant leur matrice de transition par la matrice nulle d'ordre correspondant et en définissant dans \mathcal{Q}_i les variances résiduelles correspondantes de façon à ce qu'elles soient très grandes. Pour plus de précisions, consulter Sallàs et Harville (1981). Bien qu'il ne soit pas représenté ici dans sa forme la plus générale, on sait que le modèle d'espace d'états défini par (2.3) et (2.4) englobe, à titre de cas spéciaux, nombre de modèles de série chronologique et de modèles linéaires mixtes utilisés couramment. À titre d'exemple important, mentionnons la famille des modèles ARMMI et des modèles à coefficient de régression aléatoire. Il est également facile de structurer le modèle multinitiveaux (MMN) défini par (2.1) sous forme de modèle d'espace d'états.

Pour montrer ceci, remplaçons l'indice i par t et définissons $L_t = [X_t, Z_t]$, $\alpha_t = [\beta_t, u_t]$, $H_t = \sigma^2 Z_t^0$ et $G_t = [I_p^0, I_p^q]$ où I_p^0 et I_p^q définissent la matrice d'identité et la matrice nulle d'ordre approprié. (Les matrices Z_t et X_t sont définies sous l'équation (2.2).) Le coefficient vectoriel β_t est ajouté au vecteur d'état pour des raisons de commodité. La matrice de covariance \mathcal{Q}_t est une matrice diagonale par blocs dont les deux blocs sont 0_p^0 et $Z_t^0 \Omega Z_t^0$. L'utilisation d'une matrice nulle 0_p^0 pour la covariance de $(\beta_t - \beta_{t-1})$ garantit que les coefficients β_t ne varient pas au cours du temps, conformément à (2.1). (La représentation du MMN sous forme de modèle d'espace d'états n'est pas unique.)

Pour des matrices de covariance données $\{H_t, \mathcal{Q}_t\}$, si l'on suppose que l'on connaît β_t, L_t et G_t pour toute valeur de t , le meilleur prédicteur linéaire non biaisé (MPLNB) du vecteur d'état à t importe quel point dans le temps t , fondé sur toutes les données accumulées jusqu'à ce moment-là, est obtenu de façon commode au moyen du filtre de Kalman. Supposons que $\hat{\alpha}_{t-1}$ défini le MPLNB de α_{t-1} si l'on se fonde sur les observations recueillies jusqu'au temps $(t-1)$, avec la matrice de covariance $P_{t-1} = \text{Cov}(\hat{\alpha}_{t-1} - \alpha_{t-1})$. Le MPLNB de α_t au temps $(t-1)$ est alors $\hat{\alpha}_{t-1} = G_t' \hat{\alpha}_{t-1}$ avec la matrice de covariance $P_{t-1} = \text{Cov}(\hat{\alpha}_{t-1} - \alpha_t) = G_t' P_{t-1} G_t + \mathcal{Q}_t$. Lorsque l'on recueille de nouvelles observations y_t , on met à jour le prédicteur $\hat{\alpha}_{t-1}$ et la matrice de covariance correspondante comme suit

$$\begin{aligned} \hat{\alpha}_t &= \hat{\alpha}_{t-1} + P_{t-1}^{-1} L_t' (y_t - \hat{\alpha}_{t-1}' L_t')^{-1} (y_t - \hat{\alpha}_{t-1}' L_t') \\ P_t &= P_{t-1} - P_{t-1}^{-1} L_t' F_t^{-1} L_t P_{t-1} \end{aligned} \quad (2.5)$$

où $F_t = L_t' P_{t-1}^{-1} L_t + H_t = \text{Var}(y_t - \hat{y}_{t-1}^{t-1})$ avec $\hat{y}_{t-1}^{t-1} = X_t \hat{\beta}_t + L_t \hat{\alpha}_{t-1}$ qui définit le MPLNB de y_t au temps $(t-1)$. L'application effective du filtre de Kalman nécessite l'initialisation appropriée de $\hat{\alpha}_{t-1}$ et P_{t-1} , laquelle dépend du modèle étudié. Consulter la section 4 pour l'initialisation dans les conditions du modèle proposé ici.

Les paramètres inconnus du modèle $(\beta_i, \text{éléments de } H_i, \mathcal{Q}_i \text{ et } G_i)$ sont ordinairement

$$\log(L) = -\frac{1}{2} \sum_{i=1}^n \log \{T_n^i \log(2\pi) + \frac{2}{T_n^i} \log |F_i|\} + \frac{1}{2} (X_i' - \hat{y}_{i-1}^{i-1})' F_i^{-1} (X_i' - \hat{y}_{i-1}^{i-1}). \quad (2.6)$$

estimées par EMV en se servant de façon commode de la «décomposition de l'erreur de prévision» pour établir la fonction de vraisemblance. Si l'on suppose que $\dim(y_i) = n$, le logarithme du rapport de vraisemblance prend la forme générale

$$y_{ijt} = x_{ijt}' \beta_i + z_{ijt}' v_i + z_{ijt}' u_{it} + e_{ijt} \quad (3.1)$$

À la présente section, nous proposons un modèle multinitiveaux de série chronologique qui combine des modèles transversaux à deux niveaux distincts grâce à la modélisation de l'évolution des effets aléatoires de premier et de deuxième niveaux au cours du temps. Représentant par S_i l'échantillon disponible au temps t , composé de m_i unités h de niveau 2 qui contiennent chacune n_h unités de niveau 1. La formulation de l'échantillon global en ce qui concerne les sous-ensembles S_i couvre les situations où les observations longitudinales sont recueillies à différentes périodes. Le modèle proposé tient également compte des plans de renouvellement mentionnés antérieurement et de la non-réponse lors d'un cycle. On notera qu'habituellement, les échantillons observés à différents points dans le temps ne sont pas disjoints et que l'hypothèse selon laquelle n_h est constant au cours du temps n'est pas restrictive. Pfeffermann et Nathan (article à paraître) considèrent le cas de données temporelles manquantes pour lequel cette hypothèse n'est pas vérifiée. À condition que les données manquantes le soient de façon entièrement aléatoire, il est simple de généraliser la présente méthode à ce cas. Nous supposons que le modèle qui suit est valable pour l'échantillon S_i :

où y_{ijt} est le résultat pour l'unité j de premier niveau dans l'unité h de deuxième niveau, x_{ijt} et z_{ijt} sont des vecteurs corrélés fixes connus des dimensions p et q , respectivement, v_i et u_{it} sont des coefficients vectoriels fixes (incommuns) et e_{ijt} sont des effets aléatoires indépendants de deuxième et de premier niveaux. Pour le temps t , le modèle défini par (3.1) est fondamentalement le même que le MMN défini par (2.1), mis à part le fait que nous supposons que $z_{ijt} = z_{ijt}^{m_i}$ pour toutes les valeurs de j et de t , donc que nous faisons la distinction entre les covariables de premier

L'application de modèles autorégressifs continus à des données longitudinales recueillies à intervalles irréguliers est également préconisée par Belcher, Hampton et Tunnicliffe (1984) qui recourent à des équations diffé-rentielles linéaires stochastiques pour décrire le processus de génération des données. Bryant et Day (1991, quant à eux, proposent un modèle empirique de Bayes pour l'analyse simultanée d'un système de modèles linéaires mixtes, présentant des effets aléatoires enchaînés et corrélés en série. Chi et Reinsel (1989) envisagent un test de caractérisation pour évaluer l'autocorrélation entre les erreurs individuelles dans le cas d'un modèle à effets aléatoires « conditionnellement indépendants ». Les auteurs déduisent une méthode d'estimation du maximum de vraisemblance et se servent des estimateurs pour prédire les effets aléatoires par application du modèle empirique de Bayes.

Diggle, Liang et Zeger (1994) proposent l'utilisation de modèles linéaires généralisés pour l'analyse des données longitudinales. Ils étudient un modèle de transition (Markov) où les valeurs antérieures sont considérées comme des variables prédictives supplémentaires. Ils étendent aux transitions la méthode des moindres carrés généralisés (MCG) pour estimer le maximum de vraisem-blance dans le cas de fonctions d'enchaînement linéaires, mais utilisent des fonctions conditionnelles de caractérisa-tion pour l'estimation dans le cas de fonctions d'enchaîne-ment non linéaires. Lawless (1999) s'appuie sur la chrono-logie des événements pour analyser les données longitudi-nales. Selon cette méthode, la variable dépendante est le nombre de manifestations d'un événement particulier jusqu'à un point particulier dans le temps t_i , et les proba-bilités de transition auxquelles elle est subordonnée sont modélisées sous forme de fonction des données chrono-logiques antérieures et des covariables. Zimmerman et Nunez-Anton (1997) proposent d'analyser les données longitudinales au moyen d'un modèle structure d'anté-dépendance, principalement dans le contexte de l'analyse de croissance. Aucune étude susmentionnée ne tient compte de la structure hiérarchique de la population ni du plan de sondage complexe de l'enquête.

Enfin, Skinner et Holmes (1999) envisagent, pour l'étude des données longitudinales, un modèle qui englobe un effet aléatoire « permanent » au niveau individuel et des effets aléatoires transitoires autocorrélés correspondant aux divers cycles de l'enquête. Ils étudient deux méthodes d'estimation des paramètres inconnus du modèle qui tien-nent compte, l'une et l'autre, des effets d'échantillonnage et des phénomènes « non informatifs » d'érosion. Dans la première méthode, ils considèrent les observations répétées comme des résultats multivariés corrélés et ils calculent les estimateurs à pondération probabiliste qui tiennent compte de la structure des corrélations. Dans la deuxième, ils con-sidèrent un modèle à deux niveaux où les « personnes » sont les unités de deuxième niveau et les mesures répétées, les unités de premier niveau. Aux termes de cette méthode, ils

estiment les paramètres inconnus par la méthode des moindres carrés généralisés itérés à pondération probabi-liste (MCGIP) de Pfeffermann et coll. (1998, voir la section 2.2).

2. MÉTHODES STATISTIQUES SUR LESQUELLES S'APPUIE LA MÉTHODE PROPOSÉE

2.1 Modèles multinationaux

Nous considérons ici un modèle à deux niveaux pour la variable réponse y dans une population comprenant $i = 1, \dots, M$ unités de deuxième niveau (ménages, écoles, ...) et $j = 1, \dots, N_i$ personnes dans l'unité de deuxième niveau i . Le modèle prend la forme:

$$y_{ij} = x_{ij}'\beta + z_{ij}'u_i + z_{0ij}'e_{ij}, \quad i = 1, \dots, M, \quad j = 1, \dots, N_i, \quad (2.1)$$

où x_{ij} , z_{ij} et z_{0ij} sont des valeurs connues des covariables des dimensions p , q et l respectivement, β est un vecteur paramétrique fixe de dimension p , et $u_i \sim N(0, \Omega)$ et $e_{ij} \sim N(0, \sigma^2)$ représentent les effets aléatoires indé-pendants de deuxième niveau et les résidus de premier niveau d'ordres p et l , respectivement.

L'inclusion des multiplicateurs z_{0ij} tient compte de l'hétéroscédasticité de premier niveau, tandis que les effets communs de deuxième niveau u_i expliquent les corrélations (interclasses) entre les mesures individuelles correspondant à une même unité de deuxième niveau. Dans le cas simple du « modèle à ordonnée à l'origine aléatoire », $y_{ij} = x_{ij}'\beta + u_i + e_{ij}$, ces corrélations prennent la forme bien connue $\text{Cov}(y_{ij}, y_{ik}) = \sigma_u^2 / (\sigma_u^2 + \sigma^2)$. Le modèle à coordonnée à l'origine aléatoire est souvent appliqué à l'estimation par petite région (voir plus loin).

Comme on l'a indiqué dans l'introduction, les modèles tels que (2.1) sont très souvent utilisés par les spécialistes des sciences sociales pour étudier les effets des covariables et l'interdépendance entre les observations correspondant à une même unité de niveau supérieur. Dans de tels cas, on cherche avant tout à estimer le vecteur de coefficients β et le vecteur θ des éléments distincts de Ω et σ^2 . Une autre application bien connue du modèle à deux niveaux est son utilisation pour l'estimation par petite région où les unités de deuxième niveau sont les régions géographiques ou d'autres domaines observés. Dans le cas de l'estimation par petite région, l'analyse vise à prédire la moyenne des unités de deuxième niveau (régions) $X_i'\beta + \sum_j u_j$, où X_i' et Z_i' représentent les moyennes réelles des régions corrélées et l'estimation des paramètres du modèle n'est qu'une étape intermédiaire. Pour une revue récente, consulter Rao (1999).

Le recours à l'algorithme des moindres carrés généra-lisés itérés (MCGI) est le moyen le plus commode d'estimer les paramètres inconnus du modèle (Goldstein 1986, 1995). Pour un échantillon aléatoire contenant m unités de deux-ième niveau et n_j unités de premier niveau dans chaque

On se sert de modèles simples d'espace d'états des séries chronologiques pour regrouper les modèles multivariés applicables à divers points dans le temps grâce à un ensemble d'équations linéaires de transition qui tiennent compte des relations entre les coefficients aléatoires des covariables et les effets aléatoires de niveau plus élevé des séries chronologiques. On se sert du filtre de Kalman pour estimer les paramètres du modèle et prédire les effets aléatoires pour les points courants et futurs dans le temps. On peut aussi recourir à des algorithmes de lissage pour réviser les prévisions antérieures (Harvey 1989). On applique des méthodes d'ajustement de modèle dans les conditions d'échantillonnage informatif afin de neutraliser les effets liés au processus de sélection de l'échantillon. Ces méthodes ont été étudiées ces dernières années dans le contexte de l'inférence analytique à partir d'enquêtes par sondage complexes, principalement pour l'analyse transversale de modèle à un seul niveau (Skinner et coll. 1989). Dans le présent article, nous utilisons la méthode de pondération d'échantillon applicable à la modélisation multivariée mise au point par Pfeffermann et coll. (1998). Par conséquent, les objectifs consistent ici à mettre au point des modèles et des méthodes d'estimation applicables à l'analyse longitudinale de données hiérarchiquement structurées, en tenant compte de l'ingérence des probabilités de sélection des unités d'échantillonnage. La caractéristique principale de notre méthode est que le modèle est ajusté au niveau individuel, mais qu'il contient des effets aléatoires communs de niveau plus élevé qui évoluent stochastiquement au cours du temps. Le modèle permet de prédire les effets aléatoires aux niveaux supérieur et inférieur (par exemple, ici, les effets propres aux ménages et aux personnes prises individuellement), au moyen de données pour toutes les périodes d'observation. Cette méthode devrait faciliter l'inférence axée sur un modèle à partir de données d'enquêtes complexes, puisqu'elle permet de mieux comprendre la structure et les corrélations des mesures longitudinales. Plus précisément, elle doit produire de meilleures prédictions des mesures individuelles que les modèles de séries chronologiques agrégées, qui, de par leur nature, ne permettent pas de faire la distinction entre les effets individuels (personne) et les effets communs de plus haut niveau (ménage). Ces avantages sont illustrés en partie par l'exemple de la section 6 et, de façon plus détaillée, dans un article connexe de Pfeffermann et Nathan (à paraître) qui met l'accent sur l'imputation des données manquantes. À cet égard, il est important de souligner que, même si les enregistrements longitudinaux individuels sont souvent très courts (quatre mesures par personne dans notre application), le nombre d'enregistrements est habituellement suffisamment grand pour justifier l'application des méthodes classiques d'estimation des séries chronologiques et de modélisation diagnostique. Nous ne considérons ici l'estimation des paramètres que dans le cas d'un modèle particulier, mais l'utilisation de variables à tester et de méthodes diagnostiques qui s'appuient sur les innovations

empiriques pour déterminer les modèles peut se faire dans la foulée, avec des modifications mineures, grâce à l'utilisation des méthodes d'estimation du maximum de vraisemblance et à la convergence des estimateurs paramétriques. À la deuxième section, nous passons en revue les caractéristiques principales des méthodes statistiques susmentionnées que nous appliquons dans les sections ultérieures. À la troisième section, nous proposons un modèle qui tient compte des aspects longitudinaux dont il a été question plus haut. Nous discutons des méthodes d'estimation à la quatrième section. La cinquième section contient les résultats d'une étude en simulation exécutée pour évaluer les propriétés de divers estimateurs dans le cas de divers scénarios d'échantillonnage. À la sixième section, nous présentons les résultats obtenus lors de l'ajustement du modèle à des données réelles recueillies dans le cadre de l'EPA israélienne et, enfin, à la septième section, nous résumons brièvement les extensions et les applications éventuelles du modèle.

1.2 Revue des données bibliographiques

Les travaux réalisés antérieurement dans le domaine qui nous occupe traitent principalement des données longitudinales ne provenant pas d'enquête et ne se rapportant pas à des populations hiérarchiquement structurées. Plus précisément, les auteurs des études que nous avons examinées ne considèrent aucuns l'évolution des effets de deuxième niveau (effets communs des ménages dans notre application) au cours du temps. Par exemple, Goldstein, Healy et Rasbash (1994) considèrent l'analyse de mesures répétées au moyen d'un modèle à deux niveaux où les personnes représentent le deuxième niveau et les mesures répétées, le premier niveau. Le modèle est une extension du modèle type à deux niveaux où il est tenu compte de la corrélation des mesures de premier niveau au fil du temps. Les auteurs envisagent plusieurs possibilités de modélisation de la structure d'autocorrélation, dont les modèles autoregressifs (AR) quand les mesures sont faites à intervalles réguliers et les fonctions d'autocorrélation quand les observations sont faites à intervalles inégaux. Dans ce dernier cas, ils linéarisent la fonction d'autocorrélation aux fins de l'estimation.

Plusieurs auteurs étudient l'application des modèles de série chronologique à l'analyse des données longitudinales. Dans une série d'articles publiés par Jones et ses collaborateurs (Jones et Ackerson 1990, Jones et Boadi-Boating 1991, Jones et Vecchia 1993) et dans l'ouvrage de Jones (1993), les observations étudiées sont faites à intervalles inégaux. Pour que les observations concernant un même sujet puissent être corrigées en série, les auteurs postulent des modèles autoregressifs continus à moyennes mobiles. Ces modèles présentent des effets fixes et aléatoires, mais ne tiennent pas compte de la structure hiérarchique de la population. Pour calculer la fonction de vraisemblance, on recourt à un modèle des moindres carrés pondérés et à un modèle d'espace d'états conjugués à un filtre de Kalman.

de plus en plus fréquemment pour l'analyse longitudinale à court terme, comme l'estimation des mouvements bruts entre les situations sur le marché du travail ou l'étude de la mobilité sociale. Néanmoins, l'exercice n'est pas toujours simple, étant donné la complexité des plans de sondage et les difficultés d'appariement.

2. Les enquêtes par panel à moyen terme, comme la

Survey of Income and Programme Participation (SIPP, Herriot et Kasprzyk 1984) et la *U.S. Panel Study of Income Dynamics* (PSID, *Survey Research Center*, 1984) aux États-Unis et l'Enquête sur la dynamique du travail et du revenu au Canada (EDTR, Webber 1994). Ces enquêtes diffèrent de celles sur la population active en ce sens qu'elles sont conçues spécialement pour l'analyse longitudinale et l'étude des caractéristiques économiques et sociales des ménages et des personnes. Par exemple, la SIPP comprend une étude approfondie, sous forme d'un interview rétrospectif complet tous les quatre mois. Elle fournit les antécédents professionnels complets pour la période couverte par l'enquête (de 30 à 48 mois) grâce à la combinaison d'une collecte rétrospective continue de données de remémorisation sur quatre mois et d'un rapprochement de données fournies pour des périodes plus longues.

3. Les études longitudinales par cohorte consistent à

suivre les membres d'une cohorte pendant une longue période. Par exemple, dans le cas de la *British Household Panel Survey*, qui au départ a été réalisée auprès d'un échantillon de ménages dont les adresses ont été sélectionnées en 1991, on a recueilli des données sur les mêmes ménages lors de cycles annuels subséquents pendant plus de sept ans. L'enquête permet de recueillir un large éventail de données sur les caractéristiques de la population active, les ressources économiques, la santé et le niveau de scolarité, en mettant l'accent sur les aspects longitudinaux. Durant cette enquête, on a suivi tous les membres des ménages sélectionnés au départ et completé l'échantillon par ajout d'entrants à l'échantillon de ménage, y compris les enfants mis au monde par des membres des ménages échantillonnés. D'autres études longitudinales par cohorte, comme la *British National Child Development Study* et la *British Cohort Study*, ont permis de suivre une cohorte de nouveau-nés pendant une période allant jusqu'à 40 ans. Consulter Nathan (1999) pour la description et une discussion de ces trois dernières enquêtes.

La plupart des études fondées sur les données de ces enquêtes nécessitent l'analyse longitudinale de données sur des populations qui ont une structure hiérarchique complexe, recueillies au moyen de plans de sondages complexes. Fréquemment, l'analyse type des données d'enquête longitudinale ne tient pas compte de la nature complexe du plan d'échantillonnage des données sur les données de ces trois dernières enquêtes.

Dans le présent article, nous proposons de regrouper trois méthodes statistiques distinctes pour résoudre les problèmes liés à la nature hiérarchique de la population cible, l'aspect longitudinal de l'analyse et aux effets des plans de sondage complexes. Ces trois méthodes sont la modélisation multi-niveaux (MMN), la modélisation des séries chronologiques et les méthodes d'analyse dans des conditions d'échantillonnage informatif complexe. On se sert des modèles multi-niveaux (MMN) pour tenir compte de la structure hiérarchique de nombreuses populations humaines, comme les personnes dans les ménages, les élèves dans les classes, les classes dans l'école et ainsi de suite. Ces modèles, utilisés à grande échelle par les spécialistes des sciences sociales, particulièrement dans le domaine de l'éducation, tiennent compte des effets des covariables observés aux niveaux inférieur et supérieur de la structure, à des coefficients fixes ou aléatoires. Les effets aléatoires communs inobservables qui se manifestent aux niveaux les plus élevés rendent compte d'autres variations inexplicables. La méthode des moindres carrés généralisés (MCGI) est couramment utilisée pour estimer les paramètres du modèle (Goldstein 1986, 1995).

Modélisation multiniveaux des données longitudinales d'enquêtes complexes à effets aléatoires variables en fonction du temps

MOSHE FEDER, GAD NATHAN et DANNY PFEFFERMANN¹

RÉSUMÉ

Par observation longitudinale, on entend la mesure répétée d'une même unité lors de plusieurs cycles d'enquête réalisés à intervalle fixe ou variable. On peut donc considérer chaque vecteur d'observations comme une série chronologique, couvrant habituellement une courte période. L'analyse des données recueillies sur toutes les unités permet d'ajuster des modèles de série chronologique d'ordre faible, malgré le peu de longueur des séries individuelles. Nous illustrons ce paradigme au moyen de données simulées qui imitent le plan de renouvellement de l'Enquête sur la population active (EPA) d'Israël. Conformément au plan de sondage avec renouvellement de panel de l'enquête, toute unité d'échantillonnage fait partie de l'échantillon pendant deux trimestres, en est exclue pendant deux trimestres, puis en fait de nouveau partie pendant deux trimestres. Le modèle est formé de modèles linéaires à deux niveaux établis pour des points particuliers dans le temps que l'on relie en permettant aux effets de deuxième niveau (qui correspondent aux ménages) et aux résidus de premier niveau (qui correspondent aux personnes) d'évoluer de façon stochastique au cours du temps. La vraisemblance du modèle s'établit facilement en se servant des séries chronologiques du modèle combiné. Cependant, étant donné le grand nombre de paramètres inconnus, la maximisation directe de la vraisemblance pourrait produire des estimateurs instables. Par conséquent, on adopte une méthode à deux étapes. À la première, on ajuste un modèle à deux niveaux pour chaque point dans le temps, de façon à obtenir des estimateurs pour les effets fixes et les variances. À la deuxième, on maximise la vraisemblance de la série chronologique uniquement en ce qui concerne les paramètres du modèle de la série chronologique. Cette méthode à deux étapes permet en outre de procéder à la pondération appropriée de premier et de deuxième niveau pour tenir compte des effets éventuels d'échantillonnage informatif. Les résultats empiriques de l'ajustement du modèle aux données recueillies dans le cadre de l'EPA israélienne sont également présentés.

MOTS CLÉS : Échantillonnage informatif; moindres carrés généralisés itérés (MGLI) à pondération probabiliste; plan de sondage avec renouvellement de panel; modèles d'espace d'états.

1. INTRODUCTION

1.1 Contexte et objectifs

L'ajustement de modèles aux données recueillies dans le cadre d'enquêtes longitudinales à plan de sondage complexe a suscité un intérêt croissant ces dernières années. Ce phénomène tient au fait que les décideurs et les spécialistes des sciences sociales demandent de plus en plus que l'on étudie de façon approfondie l'évolution des processus sociaux au cours du temps plutôt que de réaliser des analyses transversales qui fournissent un «instantané» à un point particulier dans le temps. L'estimation des mouvements bruts entre les états sociaux et démographiques, comme la situation d'emploi ou l'état de santé et le niveau de scolarité est un exemple bien connu. Pour une étude de ces questions et des problèmes que posent la conception d'enquête longitudinale et l'analyse des données de ces enquêtes, consulter Duncan et Kalton (1987) et Binder (1998).

Les exemples d'enquêtes que nous examinerons dans le présent article rentrent dans trois catégories:

1. Les enquêtes avec renouvellement de panel, comme les enquêtes sur la population active réalisées dans de nombreux pays. Nombre de ces enquêtes ont été

conçues au départ en vue de procéder à une analyse transversale des données sur les ménages et sur les personnes, de façon à étudier la population active et d'autres caractéristiques socioéconomiques courantes. Plus tard, des plans de sondage complexes avec renouvellement de panel ont été adoptés afin d'améliorer les comparaisons au cours du temps. Par exemple, l'Enquête sur la population active (EPA) israélienne s'appuie sur un plan de sondage avec renouvellement de panel conformément auquel chaque unité d'échantillonnage est interrogée pendant deux trimestres consécutifs, puis est exclue de l'échantillon pendant les deux trimestres suivants et est ensuite de nouveau interrogée pendant deux trimestres consécutifs. Aux États-Unis et au Brésil, on a adopté un plan de sondage plus compliqué en vertu duquel les unités d'échantillonnage sont incluses pendant huit mois, puis y sont de nouveau incluses pendant quatre mois. En Australie, au Canada et au Royaume-Uni, le plan de sondage prévoit l'observation des unités d'échantillonnage pendant plusieurs mois ou trimestres successifs, puis leur élimination de l'échantillon. Ces catégories d'enquêtes sont utilisées

- BURGESS, R.D. (1988). Évaluation des estimations du sous-dénombrement obtenues par la contre-vérification des dossiers du recensement du Canada. *Techniques d'enquête*, 14, 147-167.
- CARTER, R.G. (1990). The Measurement of net coverage error in Canadian censuses. *Recueil: Symposium 90, Mesure et amélioration de la qualité des données*, Statistique Canada.
- FELLEGLI, I.P. (1969). A theory of record linkage. *Journal of the American Statistical Association*, 64(328), 1183-1210.
- GOSSSELIN, J.-F. (1976). The methodology of the 1971 Reverse Record Check. *Survey Methodology*, 2, 180-193.
- ROMANUC, A. (1988). Une approche démographique à l'évaluation du recensement de 1986 et des estimations de population pour le Canada. *Techniques d'enquête*, 14, 169-185.
- ROYCE, D. (1993). Evaluation of the May 1993 Revised Results of the 1991 Census Coverage Studies. Division des méthodes d'enquêtes sociales, document de travail, Statistique Canada, Ottawa, Ontario.
- ROYCE, D., GERMAIN, M.-F., JULIEN, C., DICK, P., SWITZER, K., et ALLARD, B. (1994). *Couverture: Rapports techniques du recensement de 1991*. No. 92-341F au catalogue, Ottawa: Statistique Canada.
- STATISTIQUE CANADA (1999). *Couverture: Rapports techniques du recensement de 1996*. No. 92-370-XPB au catalogue.
- STATISTIQUE CANADA (2000). *Statistique démographique annuelles*. No. 21-213-XPB au catalogue.
- TOURIGNY, J., CLARK C., et PROVOST, M. (1998). Evaluation of the March 1998 Preliminary Results of the 1996 Census Coverage Studies. Division des méthodes d'enquêtes sociales, document de travail, Statistique Canada, Ottawa, Ontario.
- TOURIGNY, J., BUREAU, M., et CLARK, C. (1998). Revised Direct Estimates of 1991 Census Coverage Studies. Sept 24th release. Division des méthodes d'enquêtes sociales, document de travail, Statistique Canada, Ottawa, Ontario.

documentation du véritable rapport entre les personnes recensées et les autres résultats de classification, ce qui laisserait sous-entendre une certaine erreur de classification ou un biais dû à l'absence de correction pour le débiaisage.

Une correction pour la première des deux hypothèses exerce une influence relativement mineure sur l'estimation des personnes manquantes, tous les résultats de classification subissant une inflation ou une déflation correspondant à la différence proportionnelle pour les personnes recensées. Une correction pour la deuxième hypothèse pourrait avoir un effet assez marqué puisque l'absence d'estimation du véritable rapport laisse sous-entendre que toute la différence soit attribuée à d'autres catégories.

Si la deuxième hypothèse s'applique, une correction pourrait réduire l'erreur en fin de période pour neuf des douze provinces et territoires, c'est-à-dire pour toutes les provinces dans lesquelles l'erreur en fin de période survient dans le même sens que la différence pour les personnes recensées. Par contre, si la différence est attribuée à des problèmes de représentativité de l'échantillon, une correction subéquente aurait un effet négligeable ou de légère inflation de l'erreur en fin de période dans la plupart des provinces. De plus, l'évaluation est rendue plus complexe par le fait qu'il est difficile d'établir les chiffres comparables du recensement. Une erreur risquée d'être introduite par diverses sources, y compris l'estimation des émigrants de retour fondée sur le recensement (${}^{91FR}RE_{96pp}$), une correction exagérée ou insuffisante du chevauchement des bases de sondage, l'erreur d'échantillonnage et l'erreur non due à l'échantillonnage pour l'estimation du sous-dénombrement de 1991 et de 1996, l'erreur d'échantillonnage et l'erreur non due à l'échantillonnage pour l'estimation du surdénombrement, de même que l'erreur éventuelle de classification des personnes recensées selon la province. Dans un tel contexte, il semble justifié de poursuivre la recherche sur la véritable nature des erreurs de l'estimation des personnes recensées fondée sur la CVD.

5. CONCLUSION

Les auteurs ont montré que le programme de mesure de la couverture du recensement canadien offre des renseignements supplémentaires qui ont une valeur appréciable pour l'estimation de la population. À part la possibilité d'estimer le sous-dénombrement au recensement, on peut élargir les résultats de classification tirés de ces études de façon à obtenir une autre estimation de la croissance démographique, possiblement décomposée par composante. À l'aide de la plus importante des études de couverture (la contre-vérification des dossiers de 1996), les auteurs ont présenté une nouvelle méthode qui permet une estimation indépendante de la croissance démographique pour la période inter-censitaire. La contre-vérification des dossiers permet non seulement d'établir des estimations jugées très exactes pour l'erreur de couverture au recensement, en évitant une partie du biais de corrélation qui a nui aux études postcensitaires

REMERCIEMENTS

Nous tenons à remercier R.G. Carter et P. Dick, tous deux de Statistique Canada, de même que G. Robinson, du Bureau of the Census des États-Unis, des remarques formulées au sujet d'une version antérieure du présent document. Les remarques et suggestions du rédacteur adjoint et de deux examinateurs sont également appréciées.

BIBLIOGRAPHIE

BRACKSTONE, G.J., et GOSSSELIN, J.F. (1973). *Census Evaluation Program, 1971 RRC: Methodology Report*. Statistique Canada. Ottawa, Ontario.

Tableau 4
Composantes estimées (1991-1996) compilées par la Division de la démographie et mesure discrète (détaillée) de la CVD

	T.-N.	I.-P.-É.	N.-É.	N.-B.	QUE	ONT	MAN	SASK	ALB	C.-B.	CANADA
Naissances											
Démographie	31 748	8 803	55 994	44 444	453 556	730 520	81 485	70 382	199 484	229 511	1 905 927
CVD	31 779	8 782	55 984	44 444	454 332	729 744	81 485	70 382	199 484	229 511	1 905 927
Différence	-31	22	10	0	-776	776	0	0	0	0	0
Décès											
Démographie	-19 286	-5 692	-37 677	-28 567	-252 628	-376 760	-45 858	-40 652	-75 798	-126 935	-1 009 853
CVD	-18 530	-6 913	-43 820	-29 354	-273 617	-400 047	-56 120	-40 143	-74 640	-138 433	-1 081 605
Différence	-756	1 221	6 143	787	20 989	23 287	10 250	-509	1 158	11 498	71 752
Immigration											
Démographie	3 411	771	14 489	3 359	189 905	618 869	22 004	11 282	84 130	213 506	1 161 726
CVD	3 538	820	14 058	3 614	189 905	618 870	22 129	11 157	84 130	216 892	1 165 113
Différence	-127	-49	431	-255	0	-1	-125	125	0	-3 386	-3 387
Émigration											
Démographie	-671	-206	-2 297	-2 429	-15 490	-48 609	-5 684	-2 493	-19 718	-17 834	-115 431
CVD	-2 227	-455	-7 334	-3 889	-55 766	-168 556	-10 871	-7 133	-33 689	-31 739	-321 659
Différence	1 556	249	5 037	1 460	40 276	119 947	5 187	4 640	13 971	13 905	206 228
Migration											
Démographie	-23 074	1 643	-5 288	-3 255	-51 176	-40 850	-25 336	-26 644	7 155	167 809	984
CVD	-32 767	-886	-1 479	-2 933	-49 395	-37 505	-29 765	-25 095	-10 321	191 222	1 076
Différence	9 693	2 529	-3 809	-322	-1 781	-3 345	4 429	-1 549	17 476	-23 413	-92
Résidents non permanents											
Démographie	-1 406	164	-950	-455	-23 353	-116 602	-1 630	-777	-8 267	554	-152 722
CVD	455	236	-549	-606	-13 445	-86 934	-582	144	-5 057	4 890	-101 448
Différence	-1 861	-72	-401	151	-9 908	-29 668	-1 048	-921	-3 210	-4 336	-51 274
Total											
Démographie	-9 263	5 483	24 271	13 097	300 849	766 568	24 981	11 098	186 986	466 611	1 790 681
CVD	-17 751	1 583	16 860	11 276	252 014	655 572	6 288	9 312	159 907	472 343	1 567 404
Différence	8 488	3 900	7 411	1 821	48 835	110 996	18 693	1 786	27 079	-5 731	223 277

(sauf les terr.)

Sans être décisive, la décomposition courante indique également qu'il existe d'autres composantes problématiques à part l'émigration pour ce qui est de l'explication de l'erreur en fin de période pour certaines provinces. Ainsi, les résultats indiquent que les estimations de la migration interprovinciale comportent des erreurs pour la Colombie-Britannique et Terre-Neuve compte tenu des différences observées pour ces composantes et les erreurs en fin de période correspondantes. Dans l'ensemble, l'acceptation de la CVD relativement à ces migrations plus difficiles à estimer permettrait d'expliquer non seulement une bonne partie de cette différence de croissance, mais également une partie de l'erreur en fin de période de 1996. Pour ce qui est de l'erreur en fin de période qui reste, il convient de se pencher sur la différence observée pour les personnes éventuelles des estimations postcensitaires.

4.3. Comparaison des estimations de personnes recensées

Bien que la différence pour les personnes recensées que l'on observe à l'échelle nationale soit beaucoup plus faible que la différence documentée pour ce qui est de la croissance, pour la moitié environ des provinces, cette différence est de taille comparable sinon supérieure. Quant à l'interprétation de cette constatation, il est reconnu que la

CVD n'a jamais été conçue en fonction de la population «recensée». La priorité étant accordée à la documentation du nombre de «personnes manquantes» dans le recensement, le plan d'échantillonnage de la CVD comporte une surreprésentation des «groupes difficiles à recenser» (les jeunes adultes célibataires par exemple), de même qu'une sous-représentation des personnes «faciles à recenser». Dans l'ensemble, la comparaison pour les personnes recensées confirme l'exactitude de la CVD, les différences d'une province ou d'un territoire à l'autre n'étant pas significatives. Néanmoins, les différences constatées pour quelques provinces sont préoccupantes, et se rapprochent beaucoup du seuil statistiquement significatif à 95% au Québec (différence positive), de même que du seuil statistiquement significatif en Colombie-Britannique, en Alberta et au Manitoba (différences négatives).

Quant à l'évaluation des résultats des études de couverture de 1991, deux autres hypothèses ont été formulées pour expliquer les différences observées pour les personnes recensées (Royce 1993). À l'une des extrémités, on pourrait insister pour que toute la différence (pour une province donnée) soit expliquée en fonction de la représentativité de l'échantillon de la CVD, laissant supposer une erreur d'échantillonnage ou des lacunes quelconques dans la base de sondage. À l'autre extrémité, on pourrait insister pour que toute la différence soit attribuée à l'absence de

l'équation 1 l (équation détaillée). D'autres estimations sont fournies pour i) les naissances, iii) les décès, iii) l'immigration, iv) l'émigration, v) la migration interprovinciale et vi) le changement net du nombre de résidents non permanents. Les problèmes majeurs de l'explication de l'erreur en fin de période sont évidents dans le tableau 4, en particulier pour l'émigration.

Puisque le Canada ne possède pas de système complet d'inscription à la frontière, l'émigration est clairement l'élément le plus faible du programme des estimations de population. En l'absence de renseignements immédiats sur le nombre de personnes quittant le Canada, la CVD, avec ses procédures exhaustives de dépistage, de couplage des

Tableau 2 Résultats des études de couverture relativement à l'estimation de la population (1996 – jour du recensement)

{1}		{2}		{3}		{4=1+2+3}		{5}		{6=5-4}		{7=6/4*100}	
Dénombrement de 1996 avec ajouts aléatoires	Sous-dénombrement net de 1996	Réserves	Recensement de 1996 corrigé pour la CVD	Estimation postcensitaire de 1996 (i)	Erreur en fin de période (%)	Estimation postcensitaire de 1996 (i)	Erreur en fin de période (%)	Estimation postcensitaire de 1996 (i)	Erreur en fin de période (%)	Estimation postcensitaire de 1996 (i)	Erreur en fin de période (%)	Estimation postcensitaire de 1996 (i)	Erreur en fin de période (%)

T.-N.	551 792	9 424	0	561 216	569 950	8 734	1,56	134 557	1 149	175	0,06	909 282	20 821	0	938 593	8 490	0,91
I.-P.-É.	134 557	1 149	175	135 881	135 960	79	0,06	738 133	14 225	518	0,71	909 282	20 821	0	938 593	8 490	0,91
N.-É.	909 282	20 821	0	930 103	938 593	8 490	0,91	738 133	14 225	518	0,71	909 282	20 821	0	938 593	8 490	0,91
N.-B.	738 133	14 225	518	752 876	752 859	5 383	0,71	738 133	14 225	518	0,71	909 282	20 821	0	938 593	8 490	0,91
QUÉBEC	7 138 795	116 750	12 427	7 267 972	7 362 514	94 542	1,30	7 138 795	116 750	12 427	1,30	909 282	20 821	0	938 593	8 490	0,91
ONTARIO	10 753 573	301 368	20 849	11 075 790	11 183 050	107 260	0,97	10 753 573	301 368	20 849	0,97	909 282	20 821	0	938 593	8 490	0,91
MAN.	1 113 898	18 881	315	1 133 094	1 134 393	1 299	0,11	1 113 898	18 881	315	0,11	909 282	20 821	0	938 593	8 490	0,91
SASK.	990 237	28 051	586	1 018 874	1 014 019	-4 855	-0,48	990 237	28 051	586	-0,48	909 282	20 821	0	938 593	8 490	0,91
ALB.	2 696 826	66 327	11 287	2 774 440	2 774 832	392	0,01	2 696 826	66 327	11 287	0,01	909 282	20 821	0	938 593	8 490	0,91
C.-B.	3 724 500	142 443	3 136	3 870 079	3 831 665	-38 414	-0,99	3 724 500	142 443	3 136	-0,99	909 282	20 821	0	938 593	8 490	0,91
YUKON	30 766	1 022	0	31 788	31 032	-756	-2,38	30 766	1 022	0	-2,38	909 282	20 821	0	938 593	8 490	0,91
T.N.-O.	64 402	3 024	0	67 426	66 453	-973	-1,44	64 402	3 024	0	-1,44	909 282	20 821	0	938 593	8 490	0,91
Canada	28 846 761	723 485	49 293	29 619 539	29 800 720	181 181	0,61	28 846 761	723 485	49 293	0,61	909 282	20 821	0	938 593	8 490	0,91

(i) Estimations postcensitaires du 14 mai obtenues avec des composantes définies pour les estimations intercensitaires. Estimations définitives (24 septembre 1998) du sous-dénombrement net, 1991 et 1996.

Tableau 3 Décomposition de l'erreur en fin de période

Province/Territoire	Erreur en fin de période	Estimation de la croissance Dém. et CVD	Différence entre		Erreur-type des estimations	Différence pour les pers. rec.	Erreur-type des estimations
			Erreur-type	des estimations			

T.-N.	8 734	8 634	4 889	100	5 176	-2 836	2 462
I.-P.-É.	79	2 915	2 425	1 294	9 455	4 303	7 918
N.-B.	5 383	1 080	7 793	4 303	29 310	55 050	29 310
QUÉBEC	94 542	39 492	25 493	9 135	51 300	-16 305	10 370
ONTARIO	107 260	98 125	41 212	9 135	21 618	-34 650	22 996
ALB.	392	35 042	19 067	9 135	21 618	-34 650	22 996
SASK.	-4 855	-426	9 187	-4 429	10 200	-4 429	10 200
MAN.	1 299	17 604	10 108	-16 305	10 370	-16 305	10 370
C.-B.	-38 414	747	20 518	-39 161	22 996	-39 161	22 996
YUKON	-756	N/A	N/A	-108	270	-284	464
T.N.-O.	-973	N/A	N/A	-284	58 724	-27 498	58 724
Canada	18 1181	210 408	43 951	-27 890	58 762	-27 890	58 762

La croissance implicite (Δ^1) peut se définir comme la somme i) d'une estimation de la croissance fondée sur la CVD (à l'exclusion des réserves indiennes qui refusent), ii) d'un second terme qui dépend de la décision d'estimer les réserves indiennes qui refusent à l'aide d'un modèle indépendant et iii) d'un troisième terme comportant une comparaison entre l'estimation des personnes recensées d'après la CVD et le nombre de personnes effectivement recensées en 1996.

Ce dernier terme (la différence pour ce qui est des personnes recensées) donne lieu à une interprétation intéressante, et il est considéré comme un élément essentiel de l'évaluation de la CVD (Tourigny, Bureau et Clark 1998; Royce 1993). Des différences appréciables pour ce terme laissent supposer soit des erreurs d'échantillonnage et/ou un biais, soit une erreur de classification et/ou des problèmes de tirage de l'échantillon. Pour que cette comparaison soit utile, on élimine le surdénombrement de 1996 et l'estimation des émigrants de retour, puisque ni l'un ni l'autre ne peut faire partie de l'estimation des personnes recensées. De même, puisque la CVD tire une partie de son échantillon du recensement antérieur, elle reporte inévitablement un certain surdénombrement inhérent à ses poids, qu'il faut ensuite éliminer de son estimation des personnes recensées. Ces corrections sont comprises dans le troisième terme (le troisième couple de crochets) de l'équation 21.

Il y a inflation de l'estimation des personnes recensées sous l'effet des poids associés au surdénombrement de la base de sondage du Recensement de 1991, mais une partie seulement est associée directement à cette estimation, le reste étant réparti parmi les autres résultats de classification. Par conséquent, tous les résultats de classification figurant dans les équations mentionnées ci-dessus sont aussi légèrement surestimés. Pour la présente décomposition, cette distinction mineure est laissée de côté. C'est là une autre raison, bien que mineure, pour laquelle l'estimation de la croissance fondée sur la CVD est différente de l'estimation implicite, cette dernière n'étant pas biaisée par un tel surdénombrement.

D'après ce qui précède, l'erreur en fin de période équivalait à:

$$\frac{D}{D_{91-96}} - \frac{1}{D_{91-96}} = \left[\Delta_{91-96}^D - \{ (1-\delta) \} \right] \{ \{ I_{RR} - (I_{RR}^{91M} - I_{RR}^{RC91}) + (I_{RR}^{96M} - I_{RR}^{RC96}) \} \} - \{ \{ (P^{96} - EN^{96} - {}_{91FR}RE_{96PP} + {}_{rev}O_{91} - O_{96}) \} \} \quad (22)$$

Dans la décomposition de l'erreur en fin de période, le premier terme entre crochets [] souligne la différence entre l'estimation postcensitaire de la croissance et l'estimation de la croissance combinée de la CVD (englobant les réserves qui refusent, avec raffinement pour les estimations modélisées). Le deuxième terme (la différence pour ce qui est des personnes recensées) met en évidence les difficultés éventuelles des études de couverture. Théoriquement, en l'absence d'erreur d'échantillonnage et d'erreur non due à l'échantillonnage dans la CVD, ce dernier terme devrait être négligeable.

Dans toutes les provinces (sauf la Saskatchewan), la croissance estimée en fonction de documents administratifs est plus élevée que ne l'indique l'estimation fondée sur la CVD. À l'échelle nationale (sauf les territoires), cette divergence en matière de croissance (210 408) semble beaucoup plus importante pour l'explication de l'erreur en fin de période que la divergence en matière de personnes recensées (-27 498). Pour de nombreuses provinces, la différence de croissance tombe bel et bien dans les limites prévues en fonction de l'erreur d'échantillonnage, mais pour d'autres provinces il faut trouver une meilleure explication. Ainsi, la différence de croissance en Ontario est importante (98 125), représentant près de la moitié de la différence observée à l'échelle nationale. De même, Terre-Neuve, le Québec, l'Alberta et le Manitoba, pris ensemble, expliquent une bonne partie de cette différence. À titre d'indication des facteurs responsables de ces différences, le tableau 4 présente des comparaisons d'après

4.1. Résultats de la décomposition: erreur en fin de période

Le tableau 2 présente l'erreur en fin de période lorsque la population de 1991 et celle de 1996 ont été estimées. Si l'on ajoute le sous-dénombrement net aux chiffres publiés du Recensement de 1996, de même que les estimations indépendantes des réserves indiennes qui refusent, la population du Canada au jour du Recensement de 1996, corrigée pour l'erreur de couverture, est estimée à 29 619 539. Ce chiffre est appréciablement inférieur à l'estimation pour le jour du recensement préparée dans le cadre du programme des estimations postcensitaires: 29 800 720. La différence entre les deux chiffres, qui correspond à la différence mentionnée ci-dessus entre la croissance implicite et la croissance fondée sur des documents administratifs, est plus élevée que prévu d'après l'expérience passée, atteignant 181 181 (c'est-à-dire 0,61 % de la population au jour du Recensement de 1991).

Parmi les provinces et les territoires, l'erreur en fin de période est particulièrement marquée à Terre-Neuve (1,56 %), dans le Nord canadien (-2,38 % au Yukon et -1,44 % dans les T.N.-O.) et, chose assez étonnante, dans les trois provinces les plus importantes (1,30 % au Québec, 0,97 % en Ontario et -0,99 % en Colombie-Britannique). Dans les régions, on observe des erreurs en fin de période plus grandes que la moyenne nationale dans l'Est et le Centre du Canada (sauf en l.-P.-E.), tandis que les provinces de l'Ouest comportent des erreurs en fin de période inférieures à la moyenne nationale. Ce sont précisément ces erreurs que la décomposition courante cherche à évaluer et à expliquer.

Le tableau 3 présente les résultats de cette décomposition, selon i) la différence entre l'estimation de la croissance fondée sur des documents administratifs et l'estimation fondée sur la CVD (version simplifiée) et ii) la différence pour les personnes recensées. On y trouve également l'erreur-type associée aux estimations de la CVD.

4.2. Comparaison des estimations de la croissance

Si l'on combine les équations 5, 6 et 7, on peut réexprimer la croissance démographique comme suit:

$$P_T^{96} - P_T^{91} =$$

$$\begin{aligned} & B_{91-96}^{96} - {}_{91NP}D_{96}^{96} - {}_{91-96B}D_{96}^{96} - {}_{91NP}C_{96PP}^{96} \\ & + I_{91-96}^{96} + {}_{91FR}R_{96PP}^{96} - {}_{91NP}C_{96PP}^{96} - {}_{91PP}E_{96FR}^{96} - {}_{91NP}E_{96FR}^{96} - {}_{91-96B}E_{96FR}^{96} - {}_{91-96I}E_{96FR}^{96} \\ & - {}_{91NP}E_{96EX}^{96} - {}_{91EX}I_{96NP}^{96} \end{aligned} \quad (8)$$

Le terme final de l'équation (8) correspond à:

$${}_{91EX}I_{96NP}^{96} = {}_{NP}D_{96}^{96} - {}_{91NP}D_{96}^{96} + {}_{91NP}E_{96EX}^{96} + {}_{91NP}C_{96PP}^{96} + {}_{91NP}E_{96FR}^{96} \quad (9)$$

Par conséquent:

$$P_T^{96} - P_T^{91} =$$

$$\begin{aligned} & B_{91-96}^{96} - {}_{91NP}D_{96}^{96} - {}_{91-96B}D_{96}^{96} - {}_{91NP}C_{96PP}^{96} \\ & + I_{91-96}^{96} + {}_{91FR}R_{96PP}^{96} - {}_{91NP}C_{96PP}^{96} - {}_{91PP}E_{96FR}^{96} - {}_{91NP}E_{96FR}^{96} - {}_{91-96B}E_{96FR}^{96} - {}_{91-96I}E_{96FR}^{96} \end{aligned}$$

$$- {}_{91-96I}E_{96FR}^{96} - {}_{91NP}E_{96EX}^{96} + {}_{NP}D_{96}^{96} - ({}_{NP}D_{96}^{96} - {}_{91NP}D_{96}^{96}) \quad (10)$$

ou:

$$P_T^{96} - P_T^{91} =$$

$$\begin{aligned} & (B_{91-96}^{96} - ({}_{91PP}D_{96}^{96} - {}_{91-96B}D_{96}^{96} - {}_{91-96I}D_{96}^{96}) + ({}_{91-96I}D_{96}^{96} - {}_{91-96B}D_{96}^{96} \\ & - {}_{91NP}E_{96FR}^{96} + {}_{91-96B}E_{96FR}^{96} + {}_{91-96I}E_{96FR}^{96} - {}_{91FR}R_{96PP}^{96}) + \end{aligned} \quad (11)$$

Cette version élargie de l'équation (5) donne lieu à une ventilation de la croissance démographique au niveau national, et permet des comparaisons plus utiles avec des postes estimées à l'aide de documents administratifs. Tous les termes sauf ${}_{91FR}R_{96PP}^{96}$ et ${}_{NP}D_{96}^{96}$ peuvent être obtenus directement de la CVD de 1996. La lacune susmentionnée pour la base de sondage de la CVD nécessite une estimation indépendante des émigrants de retour, tandis que la nature de la base de sondage pour les RNP explique l'absence du dernier terme. La CVD ne se fonde pas sur une liste de tous les RNP arrivant au Canada au cours de la période inter-censitaire (comme c'était le cas pour les immigrants), mais

La CVD estime la croissance à l'aide d'information sur l'état des unités identifiées lors de deux dates «discrètes» au moins (au début et à la fin de la période intercensitaire). Malgré cette distinction conceptuelle mineure entre l'estimation «continue» et l'estimation «discrète», chaque terme de l'équation 11 (chaque couple de parenthèses) correspond plus ou moins à une composante distincte documentée à l'aide de dossiers administratifs. Le premier terme sert à identifier toutes les naissances intercensitaires (la somme pondérée de la base des naissances); le deuxième terme englobe les décès (résultats de classification pour la base des naissances, la base des personnes manquantes, la base du recensement et la base des immigrants); le troisième terme comprend tous les immigrants (la somme pondérée de la base des immigrants); le quatrième terme comprend les émigrants de retour; le cinquième terme correspond au gain net ou à la perte nette de RNP. Puisque le nombre de RNP demeurant au Canada en 1991 n'est pas disponible dans la CVD de 1996, pour le moment, ce dernier terme est obtenu à l'aide du dénombrement de 1991, après correction pour le sous-dénombrement. Encore une fois, il est possible d'exprimer cette équation au niveau provincial.

L'équation (11) permet une évaluation détaillée du programme des estimations postcensitaires. Ainsi, si il subsiste des différences entre les estimations fondées sur la CVD et les estimations postcensitaires, il est possible de déterminer dans quelle mesure les différences de croissance estimative sont attribuables à des différences de migration (typiquement assez difficiles à estimer dans le programme des estimations postcensitaires) et dans quelle mesure elles sont attribuables à des différences d'accroissement naturel. Bref, le tableau 1 englobe toutes les estimations de la croissance mentionnées ci-dessus, y compris la croissance implicite, la croissance fondée sur des documents administratifs et les deux autres estimations de la croissance fondées sur la CVD (équations simplifiées et élargies). Il existe de légères différences entre l'équation simplifiée et l'équation élargie – mais pas du tout de la même importance que pour les autres estimations (implicite, postcensitaire). En guise d'application des différences entre les deux estimations fondées sur la CVD, disons que l'équation simplifiée ne suppose pas la même classification détaillée que l'équation élargie; elle n'est pas aussi biaisée par le problème du chevauchement des bases de sondage mentionnées ci-dessus, et elle ne s'appuie pas sur le dénombrement des RNP du Recensement de 1991. Les différences observées pour les estimations qui restent font l'objet de la décomposition en cours.

toutes les naissances intercensitaires relevant d'une population – peu importe si les personnes nées se déplacent ou meurent – tandis que les naissances dans une équation discrète représentent toutes les personnes nées et demeurant au Canada à la fin de la période intercensitaire. Dans un tel contexte, il est possible d'élargir l'équation fondée sur la CVD de façon à obtenir des termes qui sont plus comparables à ceux figurant dans les estimations postcensitaires. Dès lors, l'estimation de la croissance démographique fondée sur la CVD peut servir à évaluer les composantes de la croissance démographique figurant dans la méthode des composantes.

Pour élargir cette équation, il convient de commencer par les naissances, exprimées encore une fois en fonction des résultats de classification possibles de la CVD. Comme il a été mentionné antérieurement, la naissance comme terme de l'équation (5) représente une partie seulement des naissances survenues au cours de la période intercensitaire. De façon plus générale, on peut exprimer toutes les naissances comme suit:

$$(6) \quad B_{91-96} = {}_{91-96}B_{96} + {}_{91-96}B_{96}D_{96} + {}_{91-96}B_{96}E_{96FR}$$

où:

B_{91-96} = toutes les naissances intercensitaires

${}_{91-96}B_{96}$ = toutes les naissances intercensitaires classées à terme parmi les unités dénombrées ou manquantes en 1996

${}_{91-96}B_{96}D_{96}$ = décès de naissances intercensitaires

${}_{91-96}B_{96}E_{96FR}$ = personnes hors de la population cible en 1996 mais nées au Canada au cours de la période intercensitaire

De même, on peut exprimer tous les immigrants comme suit:

$$I_{91-96} = {}_{91EX}I_{96PP} + {}_{91NP}C_{96PP} + {}_{91-96}D_{96} + {}_{91-96}E_{96FR} \quad (7)$$

où:

${}_{91EX}I_{96PP}$ = immigrants intercensitaires classés à terme parmi les unités dénombrées ou manquantes en 1996

${}_{91NP}C_{96PP}$ = tous les RNP en 1991 qui deviennent immigrants reçus et qui sont classés à terme parmi les unités dénombrées ou manquantes en 1996

${}_{91-96}D_{96}$ = décès d'immigrants reçus au cours de la période intercensitaire

${}_{91-96}E_{96FR}$ = émigrants parmi les immigrants intercensitaires (peu importe s'ils demeureraient ou non au Canada à titre de RNP en 1991)

On peut obtenir une approximation de immigrants de retour). On peut obtenir une approximation de la population cible du recensement de 1991 à l'aide de l'échantillon tiré des bases de sondage du recensement et des personnes manquantes – avec identification des résultats de classification pertinents. On peut obtenir une approximation de la population cible du recensement de 1996 à l'aide de toutes les personnes classées parmi les unités dénombrées ou manquantes en 1996. On peut obtenir le terme final (les émigrants de retour) indépendamment de la CVD à l'aide de la variable de mobilité quinquennale du Recensement de 1996, en identifiant toutes les personnes qui étaient à l'extérieur du pays il y a cinq ans (à l'exclusion des immigrants récents et des RNP). On peut exprimer cette même estimation de la croissance démographique, fondée sur la CVD, au niveau provincial en incorporant une estimation de la migration interprovinciale. Puisque la CVD se fonde sur des fichiers de soins de santé pour les deux territoires nordiques du Canada (le Yukon et les TN-O) avec des listes administratives d'adresses courantes pour le recensement en question, cette estimation de la croissance n'est pas possible pour les populations relativement peu nombreuses du Grand Nord canadien.

Il subsiste dans la CVD un problème mineur qui risque d'introduire un faible biais dans les résultats de classification. Malheureusement, il n'est pas possible d'identifier tous les RNP dans l'échantillon de la CVD, d'où le risque d'un chevauchement inconnu des bases de sondage (entre les bases de sondage du recensement, des RNP et des immigrants). Puisque les RNP, dans le recensement, ne peuvent être identifiées qu'à l'aide du questionnaire détaillé du recensement, qui est distribué à 20% environ des ménages, il est possible que quelques RNP demeurant au Canada en 1991 et sélectionnés dans la base de sondage du recensement ont également été choisis soit dans la base des immigrants, soit dans celle des RNP sans avoir été identifiés comme tels. Bien que l'on tente de corriger la CVD pour un tel chevauchement, en identifiant toute personne de ce genre dans les bases des immigrants et des RNP, il existe un biais inconnu dans la mesure où cette démarche est infructueuse. Cette difficulté d'élucidation du chevauchement entraîne la possibilité d'un nombre trop élevé d'immigrants et/ou de RNP dans l'échantillon, ou d'un nombre trop faible, si l'on élimine trop de personnes des bases de sondage mentionnées. Ce dernier résultat risque de provoquer une déflation subséquente de l'estimation de la croissance, démographique et du sous-dénombrement brut (parmi d'autres résultats de classification), tandis que la première possibilité entraîne le résultat contraire.

2.3.2. Estimation de la croissance fondée sur la CVD: une décomposition plus détaillée

Même si les estimations de la croissance démographique tant postcensitaire que fondée sur la CVD devraient être très comparables, les termes particuliers au sein de chacune ne sont pas censés être directement équivalents. Ainsi, les naissances dans les estimations postcensitaires représentent

échantillon pouvant représenter la même population cible que le recensement qu'il s'agit d'évaluer. Cette base de sondage, obtenue tout à fait indépendamment du recensement antérieur, des naissances inscrites au cours de la période intercensitaire, des listes administratives d'immigrants intercénsitaires et d'une liste à jour des résidents non permanents. Les personnes manquantes au cours du recensement antérieur sont représentées par un échantillon de cas classes comme «manquantes» lors de la CVD antérieure, en l'absence d'une liste complète de ces personnes.

Compte tenu de cet échantillon, la CVD cerne toutes les personnes qui auraient pu faire partie de l'univers du Recensement de 1996. Sauf pour une très faible sous-population d'émigrants de retour (citoyens canadiens et immigrants reçus qui étaient à l'étranger lors du recensement antérieur), l'échantillon de la CVD est complet et tout à fait représentatif. La classification subséquente (manquant, dénombré, émigré, à l'étranger, décédé ou hors du champ d'observation) est appliquée à l'estimation des personnes «manquantes» du recensement en cours. Cette classification offre également la possibilité d'autres inférences, c'est-à-dire d'une autre estimation de la croissance démographique pour la période intercensitaire.

Au moment d'estimer la croissance démographique à l'aide de la CVD, il est utile de considérer les deux équations ci-dessous. Dans la première, la population cible du Recensement de 1991 (P_T^{91}) est exprimée en fonction de tous les résultats possibles de la classification de 1996. Dans la deuxième équation, on peut procéder dans le sens contraire, c'est-à-dire exprimer la population cible du recensement de 1996 (P_T^{96}) en fonction de toutes les situations possibles pour 1991 (ou encore, dans le cas des naissances et des immigrants, pour la période intercensitaire).

$$P_T^{91} = {}^{91}P_T^{96} + {}^{91}NP^{96} + {}^{91}NP^{96}C^{96PP} + {}^{91}PP^{96}D^{96}$$

$${}^{91}NP^{96}D^{96} + {}^{91}PP^{96}E^{96FR} + {}^{91}NP^{96}E^{96EX} + {}^{91}NP^{96}C^{96PP} + {}^{91}NP^{96}B^{96} + {}^{91}PP^{96}P^{96}$$

$${}^{91}EX^{96}I^{96PP} + {}^{91}EX^{96}I^{96NP} + {}^{91}FR^{96}R^{96PP}$$

où :

${}^{91}P_T^{96}$ - les citoyens canadiens et les immigrants reçus au Canada en 1991, également ciblés par le Recensement de 1996

${}^{91}NP^{96}$ - les RNP au Canada en 1991, également ciblés par le Recensement de 1996 à titre de RNP

${}^{91}NP^{96}C^{96PP}$ - les RNP au Canada en 1991 qui sont devenus des immigrants reçus au cours de la période intercensitaire

(4)

(3)

On peut obtenir une estimation de la croissance (Δ^{RRC}) en soustrayant la première équation de la deuxième:

$$\Delta^{RRC} = {}^{91-96}B^{96} + {}^{91}EX^{96}I^{96PP} + {}^{91}EX^{96}I^{96NP} - {}^{91}PP^{96}D^{96} - {}^{91}PP^{96}E^{96FR} - {}^{91}NP^{96}E^{96EX} + {}^{91}FR^{96}R^{96PP}$$

(5)

les citoyens canadiens et les immigrants reçus au Canada en 1991 qui sont décédés au cours de la période intercensitaire

les personnes autorisées à vivre en permanence au Canada (citoyens et immigrants reçus) qui ne font pas partie de la population cible du recensement de 1996

les personnes qui n'ont jamais été citoyens ou immigrants reçus et qui ne font pas partie de la population cible du recensement en question

les RNP au Canada en 1991 qui ne sont pas devenus des immigrants reçus et qui ne font pas partie de la population cible du Recensement de 1996

les personnes nées au cours de la période 1991-1996 qui font partie de la population cible du Recensement de 1996

les personnes qui n'étaient pas au Canada en 1991, qui sont arrivées au cours de la période intercensitaire et qui sont des RNP de la population cible du Recensement de 1996

les immigrants qui ont été reçus au cours de la période intercensitaire et qui font partie de la population cible du Recensement de 1996

les immigrants de retour, c'est-à-dire les citoyens canadiens et les immigrants reçus hors de l'univers du recensement en 1991 et faisant partie de l'univers du Recensement de 1996

A l'aide des résultats de classification et des bases de sondage introduites antérieurement, il est possible d'estimer, directement à partir de la CVD de 1996 elle-même, tous les termes (exception faite pour le dernier: les

2. AUTRES ESTIMATIONS DE LA CROISSANCE DÉMOGRAPHIQUE

2.1. Estimations de la croissance fondées sur des documents administratifs: estimations postcensitaires

Le programme des estimations de population de Statistique Canada comporte l'inscription continue et l'estimation d'événements démographiques, d'après des statistiques de l'état civil et divers ensembles de données administratives. Ces événements sont additionnés ou soustraits de la population documentée lors du recensement antérieur (méthode des composantes). Dans l'estimation de la population d'une province au jour du Recensement de 1996 (P^{est96}):

$$P^{est96} = P_{91} + B_{91-96} - D_{91-96} + I_{91-96} - E_{91-96} + \Delta NPR_{91-96} + NM_{91-96} \quad (1)$$

La population de base (P_{91}) de cette estimation se fonde sur le Recensement de 1991 après correction pour tout genre d'erreurs de couverture, y compris le sous-dénombrement net au recensement mesuré à l'aide de la CVD de 1991. On peut obtenir l'estimation postcensitaire en additionnant ou en soustrayant de cette valeur de base le nombre de naissances entre les recensements (B_{91-96}), le nombre de décès (D_{91-96}), les immigrants (I_{91-96}), les émigrants (E_{91-96}), la migration interprovinciale nette (NM_{91-96}), de même que le gain net ou la perte nette de résidents non permanents (ΔNPR_{91-96}).

Les résidents non permanents (RNP) sont des personnes jouissant d'un statut juridique provisoire au Canada (titulaires d'une autorisation d'étude ou d'emploi, titulaires d'un permis ministériel, demandeurs du statut de réfugié et personnes à charge nées à l'extérieur du Canada). Contrairement à la migration interprovinciale, le gain net ou la perte nette de RNP n'est pas estimée à l'aide de données de type «flux» sur les mouvements d'entrée et de sortie de résidents non permanents, mais bien par comparaison dans le temps de données de type «stock» sur le nombre total de résidents non permanents demeurant au pays. On trouvera des renseignements supplémentaires sur la méthode, les sources et la qualité des données dans les publications trimestrielles et annuelles du programme des estimations de population (Statistique Canada 1999, 2000).

2.2. Estimation implicite de la croissance

On peut obtenir une estimation implicite de la croissance à l'aide des recensements de 1991 et de 1996, corrigés en fonction du sous-dénombrement net. Exception faite pour un petit nombre de réserves indiennes qui refusent, dont la population est estimée indépendamment, le sous-dénombrement brut a été estimé entièrement à l'aide de la CVD en 1996, tandis que le surdénombrement brut représente une

estimation combinée de trois études (la CVD, l'Étude sur les logements collectifs et l'Étude par appartement automatisé). En 1991, la CVD a été utilisée uniquement pour l'estimation du sous-dénombrement brut, tandis que le surdénombrement brut a été estimé à l'aide d'une étude plus petite, l'Étude sur les logements privés, combinée aux études par appartement automatisé et sur les logements collectifs de 1991. De plus, les personnes manquantes des réserves indiennes qui ont refusé ont été estimées dans le cadre de la CVD de 1991.

Lors de l'évaluation initiale des études de couverture de la CVD de 1996, la croissance implicite obtenue à l'aide des corrections ci-dessus a été considérée comme peu réaliste. Depuis, on a pu établir qu'une partie de l'estimation du sous-dénombrement net de 1991 était erronée, et qu'en réalité elle aurait été moins élevée si certaines améliorations méthodologiques avaient été introduites comme en 1996 (Tourigny, Clark et Provost 1998). On a pu montrer i) que le nombre de personnes d'abord classées comme manquantes en 1991 était trop élevé à cause d'une erreur de classification et ii) que l'estimation de 1991 du «surdénombrement» était trop bas. Par conséquent, les estimations de 1991 du sous-dénombrement et du surdénombrement ont été révisées de façon à refléter l'effet de ces modifications méthodologiques relativement à 1996, des estimations modélisées distinctes des réserves indiennes qui ont refusé (indépendantes de la CVD) ont été ajoutées au Recensement en 1991.

(Δ):

$$\Delta = P_{96} - P_{91} = \{P_{96}^c + U_{96} - O_{96} + IR_{96M} - IR_{96C96}\} - \{P_{91}^c + U_{91} - O_{91} + IR_{91M} - IR_{91C91}\} \quad (2)$$

où la population finale (P_{96}^c, P_{91}^c) est obtenue à l'aide de chiffres de recensement publiés antérieurement (P_{96}^c, P_{91}^c) et corrigés en fonction du sous-dénombrement (U_{96}, U_{91}) et du surdénombrement brut (O_{96}, O_{91}). Lors de l'ajout d'estimations modélisées indépendamment pour les réserves indiennes qui refusent (IR_{96M}, IR_{91M}), il est nécessaire d'éliminer la partie de l'estimation CVD du sous-dénombrement brut qui correspond à ces réserves (IR_{96C96}, IR_{91C91}). Les résultats présentés ici tiennent compte de ces changements.

2.3.1. Estimations de la croissance fondées sur la CVD

La contre-vérification des dossiers (CVD) est une procédure de couplage et d'appariement des dossiers qui sert à retrouver toutes les personnes d'un échantillon, à les interviewer et à obtenir une adresse pour le jour du recensement, tout en appariant les dossiers en fonction de documents particuliers du recensement. Cela exige la préparation d'un

Erreur de couverture au recensement: une évaluation démographique

RÉJEAN LACHAPELLE et DON KERR¹

RÉSUMÉ

Le Recensement canadien de 1996 est corrigé en fonction de l'erreur de couverture estimée surtout par la contre-vérification des dossiers (CVD). Les auteurs montrent que de nombreux renseignements supplémentaires tirés de la contre-vérification des dossiers de 1996 ont une valeur immédiate pour l'estimation de la population. En plus de rendre possible une estimation de l'erreur de couverture, les résultats de la contre-vérification des dossiers permettent d'obtenir une autre estimation de la croissance démographique, avec décomposition éventuelle par composante. Cette fonction supplémentaire de la contre-vérification des dossiers est prometteuse pour l'évaluation de l'erreur estimative de couverture au recensement et pour l'élucidation des problèmes possibles d'estimation de composantes choisies du programme des estimations de population.

MOTS CLÉS: Erreur de couverture au recensement; estimations de population; contre-vérification des dossiers.

1. INTRODUCTION

Différentes formes de contre-vérification des dossiers (CVD) sont utilisées à Statistique Canada depuis les années 1960 pour l'estimation de l'erreur de couverture du Recensement canadien (Fellegi 1969; Brackstone et Gosselin 1973; Gosselin 1976; Burgess 1988; Carter 1990; Royce, Germain, Julien, Dick, Switzer et Allard 1994; Statistique Canada 1999). À l'aide de la CVD, Statistique Canada a préparé une longue série chronologique d'estimations de population, de 1971 jusqu'à présent, qui est intégralement corrigée pour le sous-dénombrement au recensement. Les auteurs montrent ci-dessous que la CVD comporte des renseignements supplémentaires qui, d'un point de vue démographique, peuvent être utilisés à des fins d'estimation de la population.

Le programme de données démographiques de Statistique Canada fait appel à des statistiques de l'état civil, au recensement le plus récent et à différentes sources administratives en vue de la préparation d'estimations de population très exactes et à jour. Des renseignements sur les naissances, les décès, l'immigration et l'émigration, entre autres composantes démographiques, permettent d'estimer la croissance démographique depuis le recensement antérieur. Lors de chaque recensement quinquennal, un cycle prend fin et l'exacitude des estimations est mise à l'essai (Romanic 1988). On peut comparer systématiquement ces estimations de la croissance à la croissance estimative obtenue en comparant des recensements subséquents (après correction pour l'erreur de couverture au recensement).

L'interprétation de la différence observée, que l'on appelle traditionnellement l'erreur en fin de période des estimations de population intercensitaires, est loin d'être

évidente. Une erreur importante en fin de période est une indication de problèmes d'estimation de la population, dont la nature précise n'est cependant pas du tout évidente pour ce qui est de savoir quelles composantes démographiques sont responsables de l'erreur. De plus, une évaluation franche de cette erreur en fin de période risque de déceler non seulement des problèmes d'estimation de la population, mais également des lacunes éventuelles des études de couverture au recensement elles-mêmes (au début ou à la fin de la période intercensitaire).

Les auteurs décrivent la possibilité d'une autre estimation de la croissance démographique, fondée explicitement sur les résultats de classification de la CVD. Des renseignements supplémentaires facilitent grandement l'interprétation et la décomposition de cette erreur en fin de période. Trois autres estimations de la croissance démographique pour la période intercensitaire sont présentées ci-dessous, y compris la croissance estimée dans le cadre du programme d'estimations de population, la croissance implicite obtenue par comparaison de recensements consécutifs et la croissance fondée explicitement sur les résultats de classification de la CVD. La section 3 montre comment cette estimation de la CVD facilite la décomposition et l'interprétation de l'erreur en fin de période, et fournit des indications i) de biais dans des composantes choisies des estimations de population et ii) de problèmes éventuels dans les résultats de la CVD. La section 4 présente les résultats de cette décomposition et en décrit brièvement les répercussions tant pour la mesure de l'erreur de couverture au recensement que pour le programme des estimations de population.

- GHOSH, M., et RAO, J.N.K. (1994). Small area estimation: an appraisal. *Statistical Science*, 9, 55-93.
- HARVILLE, D.A. (1976). Extension of the Gauss-Markov theorem to include estimation of random effects. *Annals of Statistics*, 4, 384-395.
- HENDERSON, C.R. (1950). Estimation of genetic parameters (Résumé). *Annals of Mathematical Statistics*, 21, 309-310.
- HOGAN, H. (1992). The 1990 Post Enumeration Survey: an overview. *The American Statistician*, 46, 261-269.
- HOGAN, H. (1993). The 1990 Post Enumeration Survey: operations and results. *Journal of the American Statistical Association*, 88, 1047-1060.
- HULTING, F.T., et HARVILLE, D.A. (1991). Some Bayesian and non-Bayesian procedures for the analysis of comparative experiments and small area estimation: Computational aspects, frequentist properties, and relationships. *Journal of the American Statistical Association*, 86, 557-568.
- ISAKI, C.T., HUANG, E.T., et TSAY, J.H. (1991). Smoothing adjustment factors from the 1990 Post Enumeration Survey. *Proceedings of the Social Statistics Section, American Statistical Association*, 338-343.
- KACKAR, R.N., et HARVILLE, D.A. (1984). Approximations for standard errors of estimators for fixed and random effects in mixed models. *Journal of the American Statistical Association*, 79, 853-862.
- MORRIS, C. (1983). Parametric Empirical Bayes inference: theory and applications (avec discussion). *Journal of the American Statistical Association*, 78, 47-65.
- PEIXOTO, J.L., et HARVILLE, D.A. (1986). Comparisons of alternative predictors under the balanced one-way random model. *Journal of the American Statistical Association*, 81, 431-436.
- PRASAD, N.G.N., et RAO, J.N.K. (1990). The estimation of mean squared errors of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.
- ROBINSON, G.K. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science*, 6, 15-51.
- SEARLE, S.R. (1971). *Linear Models*. New York: Wiley.
- SINGH, M.P., GAMBINO, J., et MANTEL, H.J. (1994). Les petites régions: problèmes et solutions. *Techniques d'enquête*, 20, 3-15.
- U.S. BUREAU OF THE CENSUS. (1988). The Coverage of Population in the 1980 Census, Evaluation and Research Program, PHC(F)-4.

Nous présumons que Σ^{ee} est un multiple d'une matrice Wishart ayant d_e degrés de liberté et nous établissons approximativement $\Sigma^{\phi\phi} - \Sigma^{\phi\phi}$ avec $(1 - \phi)(\Sigma^{ee} - \Sigma^{ee})$. Nous avons l'équation suivante

$$(1 - \phi)^2 E \left\{ L \left(\Sigma^{ee} - \Sigma^{ee} \right) \Sigma^{\phi\phi} \Sigma^{\phi\phi} \right\} L' = (1 - \phi)^2 d_e^{-1} L \left[\Sigma^{\phi\phi} \Sigma^{\phi\phi} \Sigma^{\phi\phi} \Sigma^{\phi\phi} \right] L' + \Sigma^{\phi\phi} \Sigma^{\phi\phi} \Sigma^{\phi\phi} \Sigma^{\phi\phi} \Sigma^{\phi\phi} \Sigma^{\phi\phi} L' \tag{B.9}$$

Le terme dominant est celui associé à la trace, le seul que nous retenons dans notre approximation. Par conséquent, voici une approximation de la variance de β

$$V_{\beta\beta} = L \Sigma^{zz} L' + d_e^{-1} (1 - \phi)^2 \text{tr} \left\{ \Sigma^{\phi\phi} \Sigma^{\phi\phi} \Sigma^{\phi\phi} \Sigma^{\phi\phi} \right\} L \Sigma^{ee} L' \tag{B.10}$$

En combinant les résultats en (B.6), (B.7) et (B.9), nous obtenons l'estimateur brut de la variance de la variable explicative (31) suivant

$$\hat{V} \{ \hat{y}^{\phi} \} = \hat{H}^{\phi} \hat{\Sigma}^{ww} \hat{H}^{\phi} + \hat{G}^{\phi} \hat{\Sigma}^{ee} \hat{G}^{\phi} + \hat{H}^{\phi} X' \hat{V}_{\beta\beta} X \hat{H}^{\phi} + \hat{\Gamma}_{44} + \hat{\Gamma}_{33} \tag{B.11}$$

$$\hat{H}^{\phi} = I - \hat{G}^{\phi}$$

$$\hat{V}_{\beta\beta} = \hat{L}^{\phi} \hat{\Sigma}^{zz} \hat{L}^{\phi} + d_e^{-1} (1 - \phi)^2 \times \text{tr} \left\{ \hat{\Sigma}^{\phi\phi} \hat{\Sigma}^{\phi\phi} \hat{\Sigma}^{\phi\phi} \hat{\Sigma}^{\phi\phi} \right\} \hat{L}^{\phi} \hat{\Sigma}^{ee} (1 + \delta^{\phi}) \hat{L}^{\phi}$$

$$\hat{L}^{\phi} = \left(X' \hat{S}^{-1} X' \right)^{-1} X' \hat{S}^{-1} \hat{\Sigma}^{-1} \hat{\Sigma}^{\phi\phi} = \left(\hat{\Sigma}^{\phi\phi} + \delta^{\phi} \hat{\Sigma}^{\phi\phi} \right)^{-1} \hat{\Sigma}^{zz} = \hat{\Sigma}^{-1} \hat{\Sigma}^{ww} + \hat{\Sigma}^{ee}$$

$$\delta^{\phi} = \left[d_e^{-1} - \text{tr} \left\{ \hat{\Sigma}^{\phi\phi} \hat{\Sigma}^{\phi\phi} \right\}^{-1} \text{tr} \left\{ \hat{\Sigma}^{-1} \hat{\Sigma}^{ee} \right\} \right]$$

$$\hat{\Gamma}_{44} = d_e^{-1} (1 - \phi)^2 \text{tr} \left\{ \hat{\Sigma}^{\phi\phi} \hat{\Sigma}^{\phi\phi} \hat{\Sigma}^{\phi\phi} \hat{\Sigma}^{\phi\phi} \right\} \hat{G}^{\phi} \hat{\Sigma}^{ee} \hat{G}^{\phi},$$

$$\hat{\Gamma}_{33} = \hat{H}^{\phi} \begin{pmatrix} \hat{V}_{11} \hat{V} \{ \hat{\epsilon}_1^{\phi} \} & \hat{V}_{12} \hat{V} \{ \hat{\epsilon}_1^{\phi}, \hat{\epsilon}_2^{\phi} \} \\ \hat{V}_{21} \hat{V} \{ \hat{\epsilon}_2^{\phi}, \hat{\epsilon}_1^{\phi} \} & \hat{V}_{22} \hat{V} \{ \hat{\epsilon}_2^{\phi} \} \end{pmatrix} \hat{H}^{\phi},$$

$$\hat{V} = \begin{pmatrix} \hat{V}_{11} & \hat{V}_{12} \\ \hat{V}_{21} & \hat{V}_{22} \end{pmatrix} = \hat{\Sigma}^{-1} \hat{\Sigma}^{zz} \hat{\Sigma}^{-1}$$

$V \{ \hat{\epsilon}_j^{\phi} \}$, $j = 1, 2$, est la variance estimée de $\hat{\epsilon}_j^{\phi}$, et $C \{ \hat{\epsilon}_1^{\phi}, \hat{\epsilon}_2^{\phi} \}$ est la covariance estimée entre $\hat{\epsilon}_1^{\phi}$ et $\hat{\epsilon}_2^{\phi}$. Voir l'annexe A. L'estimateur de la variance de β est corrigé en fonction du fait que $(X' \hat{\Sigma}^{-1} X)^{-1}$ est un estimateur biaisé de

BIBLIOGRAPHIE

BATTESE, G.E., HARTER, R.M., et FULLER, W.A. (1988). An error components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

GRESSIE, N. (1992). Estimation du maximum de vraisemblance avec contrainte (MVC) dans le lissage des taux de sous-dénombrement du recensement selon l'approche empirique de Bayes. *Techniques d'enquête*, 18, 83-103.

EFRON, B., et MORRIS, C. (1972). Limiting the risk of Bayes and Empirical Bayes estimates - Part II: The Empirical Bayes case. *Journal of the American Statistical Association*, 67, 130-139.

ERICKSEN, E.P., et KADANE, J.B. (1985). Estimating the population in a census year (avec discussion). *Journal of the American Statistical Association*, 80, 98-131.

ERICKSEN, E.P., KADANE, J.B., et TUREY, J.W. (1989). Adjusting the 1980 Census of Population and Housing (avec discussion). *Journal of the American Statistical Association*, 84, 927-944.

FAY, R.E. (1987). Application of multivariate regression to small domain estimation. Dans *Small Area Statistics*, (eds. R. Platek, J.N.K. Rao, C.-E. Särndal et M.P. Singh). New York: Wiley, 91-102.

FAY, R.E. (1990). VPLX: Variance estimates for complex samples. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 266-271.

FAY, R.E. (1992). Inferences for Small Domain Estimates From the 1990 Post Enumeration Survey. Document non-publié, U.S. Bureau of the Census.

FAY, R.E., et HERRIOTT, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

FULLER, W.A., et HARTER, R.M. (1987). The multivariate components of variance model for small area estimation. Dans *Small Area Statistics*, (eds. R. Platek, J.N.K. Rao, C.-E. Särndal et M.P. Singh). New York: Wiley, 103-123.

GHOSH, M. (1992). Hierarchical and Empirical Bayes multivariate estimation. Dans *Current Issues in Statistical Inference: Essays in Honor of D. Basu*, (eds. M. Ghosh et P.K. Pathak), IMS Lecture Notes Monograph Series, 17, 151-177.

et la valeur de $\mathbf{H} = \mathbf{H}_0$ est définie à l'équation (31).

$$\beta - \beta = (\mathbf{X}'\Sigma_{-1}^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma_{-1}^{-1}\mathbf{z} \quad (\text{B.4})$$

Maintenant, Σ_{-1}^{-1} est indépendant de \mathbf{z} et $\mathbf{V} - \mathbf{X}\beta$ est non-corrélé avec $\beta - \beta$ si la valeur vraie de Σ_{-1}^{-1} est utilisée plutôt que la valeur de Σ_{-1}^{-1} . Par conséquent,

$$E\{\mathbf{H}_0' \mathbf{X}(\beta_0 - \beta) \mathbf{z}'(\mathbf{H}_0 - \mathbf{H}_0)\} = 0, \quad (\text{B.5})$$

où la valeur de \mathbf{H}_0 est établie au moyen de $\mathbf{Y} - \mathbf{X}\beta$ dans les estimateurs des éléments de Σ_{-1}^{-1} définie à l'annexe A et $\mathbf{H}_0 = \Sigma_{-1}^{-1}\Sigma_{-1}^{-1}$. Nous fixons la covariance entre β et \mathbf{H}_0 égale à zéro pour toutes les valeurs de ϕ . Maintenant,

$$\mathbf{H}_0 = \Sigma_{-1}^{-1}[\phi\mathbf{D}^{ee} + (1 - \phi)\Sigma_{-1}^{-1}]$$

$$= [\Sigma_{-1}^{-1}\mathbf{D}^{ee} + \phi\mathbf{D}^{ee} + (1 - \phi)\Sigma_{-1}^{-1}]\mathbf{D}^{ee} + (1 - \phi)\Sigma_{-1}^{-1}$$

et

$$\mathbf{H}_0 - \mathbf{H} = \Sigma_{-1}^{-1}[\phi(\mathbf{D}^{ee} - \mathbf{D}^{ee}) + (1 - \phi)(\Sigma_{-1}^{-1} - \Sigma_{-1}^{-1})]$$

$$- \Sigma_{-1}^{-1}[\Sigma_{-1}^{-1}\mathbf{D}^{ee} + \phi(\mathbf{D}^{ee} - \mathbf{D}^{ee})]$$

$$+ (1 - \phi)(\Sigma_{-1}^{-1} - \Sigma_{-1}^{-1})[\mathbf{H}_0]$$

$$= \Sigma_{-1}^{-1}[\phi(\mathbf{D}^{ee} - \mathbf{D}^{ee} - \mathbf{D}^{ee} + (1 - \phi)(\Sigma_{-1}^{-1} - \Sigma_{-1}^{-1}))]\mathbf{G}^{\phi}$$

$$- \Sigma_{-1}^{-1}[\Sigma_{-1}^{-1}\mathbf{D}^{ee} - \Sigma_{-1}^{-1}\mathbf{H}^{\phi}]$$

où $\mathbf{G}^{\phi} = \mathbf{I} - \mathbf{H}_0$. L'apport de $\mathbf{D}^{ee} - \mathbf{D}^{ee}$ à la variance de \mathbf{H}_0 est faible relativement à l'apport de $\Sigma_{-1}^{-1} - \Sigma_{-1}^{-1}$. Par conséquent, nous omettons $\mathbf{D}^{ee} - \mathbf{D}^{ee}$ dans notre approximation de la variance. Puis, l'espérance mathématique est

$$E\{(\mathbf{I} - \mathbf{H}_0)'(\Sigma_{-1}^{-1} - \Sigma_{-1}^{-1})\Sigma_{-1}^{-1}\mathbf{z}'\Sigma_{-1}^{-1}\mathbf{z}(\mathbf{I} - \mathbf{H}_0)\}$$

$$(\Sigma_{-1}^{-1} - \Sigma_{-1}^{-1})(\mathbf{I} - \mathbf{H}_0)\}$$

$$= d_{-1}^e \mathbf{G}^{\phi}[\Sigma_{-1}^{-1}\mathbf{D}^{ee} + \Sigma_{-1}^{-1}\mathbf{V}^{\phi}]\mathbf{G}^{\phi} \quad (\text{B.6})$$

où $\mathbf{V} = \Sigma_{-1}^{-1}\Sigma_{-1}^{-1}\Sigma_{-1}^{-1}$, parce que la valeur de \mathbf{z} est indépendante de Σ_{-1}^{-1} . Nous omettons aussi le terme $d_{-1}^e \mathbf{G}^{\phi}\Sigma_{-1}^{-1}\Sigma_{-1}^{-1}\Sigma_{-1}^{-1}\mathbf{G}^{\phi}$ dans notre approximation de la variance.

L'espérance pour le terme renfermant $(\Sigma_{-1}^{-1} - \Sigma_{-1}^{-1})$ est l'équation suivante

$$E\{\mathbf{H}'(\Sigma_{-1}^{-1} - \Sigma_{-1}^{-1})\Sigma_{-1}^{-1}\mathbf{z}'\Sigma_{-1}^{-1}(\Sigma_{-1}^{-1} - \Sigma_{-1}^{-1})\mathbf{H}\}$$

où

$$\Sigma_{-1}^{-1} - \Sigma_{-1}^{-1} = \begin{pmatrix} 0 & \mathbf{I}^{n_2}(\sigma_2^2 - \sigma_2^2) \\ \mathbf{I}^{n_1}(\sigma_1^2 - \sigma_1^2) & 0 \end{pmatrix}$$

Etablissant approximativement l'espérance en considérant la valeur de \mathbf{z} comme indépendante de Σ_{-1}^{-1} , nous obtenons

$$\mathbf{H}^{\phi} = \begin{pmatrix} \mathbf{V}_{11}'\{\sigma_1^2\} & \mathbf{V}_{12}'\{\sigma_1^2, \sigma_2^2\} \\ \mathbf{V}_{21}'\{\sigma_2^2, \sigma_1^2\} & \mathbf{V}_{22}'\{\sigma_2^2\} \end{pmatrix} \mathbf{H}^{\phi} \quad (\text{B.7})$$

où

$$\mathbf{V} = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix} = \Sigma_{-1}^{-1}\Sigma_{-1}^{-1}\Sigma_{-1}^{-1}$$

L'extension de Taylor de $\beta - \beta$ est la suivante

$$\beta - \beta = (\mathbf{X}'\Sigma_{-1}^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma_{-1}^{-1}\mathbf{z}$$

$$= (\mathbf{X}'\Sigma_{-1}^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma_{-1}^{-1}\mathbf{z}$$

$$+ (\mathbf{X}'\Sigma_{-1}^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma_{-1}^{-1}(\Sigma_{-1}^{-1} - \Sigma_{-1}^{-1})$$

$$\times \Sigma_{-1}^{-1}\mathbf{X}(\mathbf{X}'\Sigma_{-1}^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma_{-1}^{-1}\mathbf{z}$$

$$- (\mathbf{X}'\Sigma_{-1}^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma_{-1}^{-1}$$

$$\times (\Sigma_{-1}^{-1} - \Sigma_{-1}^{-1})\mathbf{z} + \text{reste.}$$

$$= (\mathbf{X}'\Sigma_{-1}^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma_{-1}^{-1}\mathbf{z}$$

$$+ \mathbf{L}(\Sigma_{-1}^{-1} - \Sigma_{-1}^{-1})\mathbf{D}^{ee}\mathbf{z}$$

$$- \mathbf{L}(\Sigma_{-1}^{-1} - \Sigma_{-1}^{-1})\mathbf{z} + \text{reste}$$

(B.8)

où $\mathbf{Q} = \mathbf{X}(\mathbf{X}'\Sigma_{-1}^{-1}\mathbf{X})^{-1}\mathbf{X}'$ et $\mathbf{L} = (\mathbf{X}'\Sigma_{-1}^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma_{-1}^{-1}$. Si $\Sigma_{-1}^{-1} = \Sigma_{-1}^{-1}$ et si Σ_{-1}^{-1} sont distribués comme un multiple d'une matrice Wishart ayant des degrés de liberté d_e , indépendants de \mathbf{z} , alors

$$E\{\mathbf{L}(\Sigma_{-1}^{-1} - \Sigma_{-1}^{-1})\mathbf{D}^{ee}\mathbf{z}'\Sigma_{-1}^{-1}\mathbf{z}(\Sigma_{-1}^{-1} - \Sigma_{-1}^{-1})\mathbf{L}'\}$$

$$\times (\Sigma_{-1}^{-1} - \Sigma_{-1}^{-1})\mathbf{L}'\}$$

$$= d_{-1}^e \mathbf{L}[\Sigma_{-1}^{-1}\mathbf{D}^{ee} + \mathbf{Q}]\mathbf{L}'$$

$$= d_{-1}^e \mathbf{L}(\mathbf{X}'\Sigma_{-1}^{-1}\mathbf{X})^{-1}\mathbf{X}'(k+1).$$

Selon une approximation semblable

$$E\{\mathbf{L}(\Sigma_{-1}^{-1} - \Sigma_{-1}^{-1})\Sigma_{-1}^{-1}\mathbf{X}\mathbf{L}\Sigma_{-1}^{-1}(\Sigma_{-1}^{-1} - \Sigma_{-1}^{-1})\mathbf{L}'\}$$

$$= E\{\mathbf{L}(\Sigma_{-1}^{-1} - \Sigma_{-1}^{-1})\Sigma_{-1}^{-1}\mathbf{X}\mathbf{L}\Sigma_{-1}^{-1}(\Sigma_{-1}^{-1} - \Sigma_{-1}^{-1})\mathbf{L}'\}$$

$$= E\{\mathbf{L}(\Sigma_{-1}^{-1} - \Sigma_{-1}^{-1})\mathbf{D}^{ee}\Sigma_{-1}^{-1}(\Sigma_{-1}^{-1} - \Sigma_{-1}^{-1})\mathbf{L}'\}$$

$$= d_{-1}^e \mathbf{L}(\mathbf{X}'\Sigma_{-1}^{-1}\mathbf{X})^{-1}(k+1).$$

En fonction de ce résultat, nous utilisons l'approximation suivante

$$\beta - \beta = (\mathbf{X}'\Sigma_{-1}^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma_{-1}^{-1}\mathbf{z} - \mathbf{L}(\Sigma_{-1}^{-1} - \Sigma_{-1}^{-1})\mathbf{z}.$$

ANNEXE A: Estimation de Σ^{ww}

Les estimateurs des valeurs de σ_1^2 et de σ_2^2 de Σ^{ww} sont

modèles sur l'analyse des estimateurs de variance. Le processus d'estimation comporte plusieurs étapes dans le cadre desquelles on utilise des estimateurs améliorés d'une étape à l'autre. Nous divisons le problème de la régression

comme suit

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} z_1 \\ z_2 \end{pmatrix},$$

où (Y_1, X_1) comprend les observations pour les groupes minoritaires et (Y_2, X_2) regroupe les autres observations. Disons que Y_1 représente les observations du vecteur-colonne dimensionnel n_1 et que Y_2 représente les observations du vecteur-colonne dimensionnel n_2 . Un estimateur initial de $(\beta_1, \beta_2)'$ est:

$$\begin{pmatrix} \tilde{\beta}_1 \\ \tilde{\beta}_2 \end{pmatrix} = \begin{pmatrix} (X_1' \tilde{\Sigma}^{-1} X_1)^{-1} X_1' \tilde{\Sigma}^{-1} Y_1 \\ (X_2' \tilde{\Sigma}^{-1} X_2)^{-1} X_2' \tilde{\Sigma}^{-1} Y_2 \end{pmatrix},$$

où

$$\tilde{\Sigma}^{ee} = \begin{pmatrix} \tilde{\Sigma}^{ee11} & \tilde{\Sigma}^{ee12} \\ \tilde{\Sigma}^{ee21} & \tilde{\Sigma}^{ee22} \end{pmatrix}$$

est divisée conformément à la partition de Y .

Les estimateurs initiaux de σ_1^2 et de σ_2^2 sont les suivants

$$\hat{\sigma}_2^2 = \max \left\{ (Y_2' - X_2' \tilde{\beta}_2)' \tilde{\Sigma}^{-1} (Y_2 - X_2' \tilde{\beta}_2) - g_{21}, 0 \right\},$$

pour $i = 1, 2$, où

$$g_{1i} = \text{tr} \left\{ \tilde{\Sigma}^{ee11} (I^{n_i} - X_i' \tilde{\Sigma}^{-1} X_i) \tilde{\Sigma}^{-1} (I^{n_i} - X_i' \tilde{\Sigma}^{-1} X_i) \tilde{\Sigma}^{ee1i} \right\},$$

$$g_{2i} = \text{tr} \left\{ (I^{n_i} - X_i' \tilde{\Sigma}^{-1} X_i) \tilde{\Sigma}^{ee1i} (I^{n_i} - X_i' \tilde{\Sigma}^{-1} X_i) \tilde{\Sigma}^{ee2i} \right\},$$

$$\tilde{A}^{N_{i1i}} = (X_i' \tilde{\Sigma}^{-1} X_i)^{-1} X_i' \tilde{\Sigma}^{-1}$$

et I^{n_i} est la matrice identité $n_i \times n_i$.

Les estimateurs finaux sont les suivants

$$\hat{\sigma}_2^2 = \max \left\{ (Y_2' - X_2' \hat{\beta}_2)' \tilde{\Sigma}^{-1} (Y_2 - X_2' \hat{\beta}_2) - g_{21}, 0 \right\},$$

pour $i = 1, 2$, où

$$g_{1i} = \text{tr} \left\{ \tilde{\Sigma}^{ee11} (I^{n_i} - X_i' \tilde{\Sigma}^{-1} X_i) \tilde{\Sigma}^{-1} (I^{n_i} - X_i' \tilde{\Sigma}^{-1} X_i) \tilde{\Sigma}^{ee1i} \right\}$$

$$g_{2i} = \text{tr} \left\{ (I^{n_i} - X_i' \tilde{\Sigma}^{-1} X_i) \tilde{\Sigma}^{ee1i} (I^{n_i} - X_i' \tilde{\Sigma}^{-1} X_i) \tilde{\Sigma}^{ee2i} \right\}$$

$$\tilde{\Sigma}^{ee11} = \tilde{\Sigma}^{ee11} + \hat{\sigma}_2^2 I^{n_1}$$

$$\tilde{\beta}_1' = (X_1' \tilde{\Sigma}^{-1} X_1)^{-1} X_1' \tilde{\Sigma}^{-1} Y_1 = \tilde{A}^{N_{111}} Y_1,$$

où

$$= e - H' H' z + H' \tilde{\Sigma}^{-1} (H' - H' \tilde{\Sigma}^{-1} H' z + O_p(n^{-1})), \quad (B.3)$$

$$\hat{y}^\phi - y = e - H' \tilde{\Sigma}^{-1} (Y - X \hat{\beta}^\phi)$$

Selon une extension de Taylor

$$y = X\beta + w \quad \text{and} \quad z = w + e.$$

le vecteur vrai inconnu à prédire et écrivons

Aux fins de l'estimation de la variance, nous présupposons que Σ^{ee} est un estimateur sans biais de Σ^{ee} distribué comme un multiple d'une matrice Wishart ayant des degrés de liberté d_e indépendants de (w, e) . Nous disons que y est

et la valeur de Σ^{ww} est définie à l'équation (30) du texte.

$$\begin{pmatrix} e \\ w \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma^{ee} & 0 \\ 0 & \Sigma^{ww} \end{pmatrix} \right), \quad (B.2)$$

où Y est un vecteur-colonne dimensionnel n , X est une

$$Y = X\beta + w + e, \quad (B.1)$$

Notre modèle est le suivant

des variables explicatives

ANNEXE B: Approximations de la variance

Voir Searle (1971, chapitre 2 et p. 435).

$$M_{22} = \begin{pmatrix} 0 & 0 \\ g_{22}^{-1} (I^{n_2} - X_2' \tilde{\Sigma}^{-1} X_2) \tilde{\Sigma}^{-1} (I^{n_2} - X_2' \tilde{\Sigma}^{-1} X_2) \tilde{\Sigma}^{ee22} & 0' \end{pmatrix}.$$

et

$$M_{11} = \begin{pmatrix} g_{11}^{-1} (I^{n_1} - X_1' \tilde{\Sigma}^{-1} X_1) \tilde{\Sigma}^{-1} (I^{n_1} - X_1' \tilde{\Sigma}^{-1} X_1) \tilde{\Sigma}^{ee11} & 0' \\ 0 & 0 \end{pmatrix}$$

où

$$\hat{C}\{\hat{\sigma}_1^2, \hat{\sigma}_2^2\} = 2 \text{tr} \left\{ \tilde{\Sigma}^{ee} M_{11} \tilde{\Sigma}^{ee} M_{22} \right\} + 2 d_e^{-1} \text{tr} \left\{ \tilde{\Sigma}^{ee} M_{11} \tilde{\Sigma}^{ee} M_{22} \right\},$$

pour $i = 1, 2$. La covariance estimée est

$$\begin{aligned} &= 2 \hat{g}_{2i}^{-2} \\ &\times \text{tr} \left\{ \left[\tilde{\Sigma}^{ee11} (I^{n_i} - X_i' \tilde{\Sigma}^{-1} X_i) \tilde{\Sigma}^{-1} (I^{n_i} - X_i' \tilde{\Sigma}^{-1} X_i) \tilde{\Sigma}^{ee1i} \right] \right. \\ &\quad \left. + 2 \hat{g}_{2i}^{-2} d_e^{-1} \right. \\ &\quad \left. \times \text{tr} \left\{ \left[\tilde{\Sigma}^{ee11} (I^{n_i} - X_i' \tilde{\Sigma}^{-1} X_i) \tilde{\Sigma}^{-1} (I^{n_i} - X_i' \tilde{\Sigma}^{-1} X_i) \tilde{\Sigma}^{ee1i} \right] \right\} \right\} \end{aligned}$$

Les estimateurs de la variance sont les suivants

Tableau 4
Estimation du pourcentage du sous-dénombrement net par groupe poststrat

Groupe poststrat	\bar{Y}	e.r. (\bar{Y})	\bar{Y}_0	e.r. (\bar{Y}_0)
Blancs non hispaniques propriétaires-grande région urbaine	-2,08	1,04	-0,63	0,60
1. Nord-Est	0,69	0,72	0,38	0,44
2. Sud	-0,26	0,39	-0,13	0,31
3. Midwest	-0,34	0,64	-0,02	0,44
4. Ouest	-1,07	0,48	-0,73	0,35
5. Nord-Est	0,52	0,43	0,53	0,33
6. Sud	-0,10	0,40	0,01	0,31
7. Midwest	0,63	0,58	0,30	0,40
8. Ouest	-0,53	0,69	-0,28	0,47
9. Nord-Est	0,18	0,69	0,58	0,45
10. Sud	-0,70	1,16	0,16	0,64
11. Midwest	0,29	0,69	0,38	0,46
12. Ouest	1,17	1,43	2,07	0,61
13. Nord-Est	2,62	1,56	3,53	0,64
14. Sud	2,39	1,70	2,53	0,60
15. Midwest	3,28	1,72	3,10	0,58
16. Ouest	3,53	1,62	2,29	0,61
17. Nord-Est	3,30	1,86	3,67	0,67
18. Sud	3,30	1,86	3,67	0,67
19. Midwest	1,24	1,13	2,39	0,53
20. Ouest	4,70	1,47	3,20	0,57
Blancs non hispaniques localitaires-région non urbaine	6,97	4,67	3,54	0,92
21. Nord-Est	6,65	1,93	3,60	0,66
22. Sud	2,93	1,60	2,36	0,66
23. Midwest	6,48	2,06	3,48	0,67
24. Ouest	1,65	1,96	0,97	0,91
25. Nord-Est	2,20	0,94	2,30	0,70
26. Sud	0,82	0,88	1,13	0,67
27. Midwest	6,49	2,16	2,54	0,96
28. Ouest	1,36	1,01	2,05	0,72
29. E-U	3,64	2,03	2,85	0,98
Noirs localitaires-grande région urbaine	9,13	1,93	5,57	0,96
31. Nord-Est	6,69	2,17	6,42	1,10
32. Sud	6,38	1,91	5,43	1,03
33. Midwest	11,06	3,35	6,04	1,12
34. Ouest	4,33	1,28	4,99	0,82
Noirs localitaires-grande région urbaine	4,84	5,95	5,90	1,24
36. E-U	0,68	4,44	3,00	1,18
37. Nord-Est	2,52	0,95	2,52	0,72
38. Sud	-4,14	2,38	0,17	0,97
39. Midwest	2,98	0,92	2,89	0,68
40. Ouest	0,95	1,70	2,32	0,87
41. E-U	2,80	2,83	2,88	1,16
Hispaniques non Noirs propriétaires-région non urbaine	7,21	4,04	5,85	1,27
43. Nord-Est	1,03	3,11	7,35	1,15
44. Sud	7,11	3,74	5,71	1,21
45. Midwest	6,29	2,09	6,45	0,98
46. Ouest	7,07	3,10	6,26	1,09
Hispaniques non Noirs localitaires-grande région urbaine	18,76	7,24	7,51	13,8
48. E-U				

Les estimations des erreurs-types des variables explicatives ont été calculées au moyen de l'approximation de la variance brute de l'annexe B. La moyenne des ratios de l'erreur-type de y_0 à y_0 pour certaines valeurs sélectionnées de ϕ figurent au tableau 3. L'ordonnement des ratios pour les 48 groupes de strates est à peu près le même que pour les 336 poststrates initiales. On constitue un groupe de poststrates en combinant les sept cellules âge-sexe au sein d'un classement donné race-mode d'occupation-degré d'urbanisation-région. En fonction de ces calculs, une valeur ϕ de 0,5 ou de 0,6 est l'estimateur préféré, bien que l'estimation des différences en matière d'efficacité ne soit pas importante. Tout membre de la catégorie ϕ est de loin supérieur à l'estimateur Y initial. L'efficacité de la variance estimée moyenne est d'environ 400% pour les variables explicatives ϕ , relativement aux estimateurs poststrates initiaux.

Les estimations des erreurs-types des variables explicatives ont été calculées au moyen de l'approximation de la variance brute de l'annexe B. La moyenne des ratios de l'erreur-type de y_0 à y_0 pour certaines valeurs sélectionnées de ϕ figurent au tableau 3. L'ordonnement des ratios pour les 48 groupes de strates est à peu près le même que pour les 336 poststrates initiales. On constitue un groupe de poststrates en combinant les sept cellules âge-sexe au sein d'un classement donné race-mode d'occupation-degré d'urbanisation-région. En fonction de ces calculs, une valeur ϕ de 0,5 ou de 0,6 est l'estimateur préféré, bien que l'estimation des différences en matière d'efficacité ne soit pas importante. Tout membre de la catégorie ϕ est de loin supérieur à l'estimateur Y initial. L'efficacité de la variance estimée moyenne est d'environ 400% pour les variables explicatives ϕ , relativement aux estimateurs poststrates initiaux.

Tableau 3

Moyenne du ratio de l'erreur-type de y_0 et de Y par rapport à l'erreur-type de y_0		et de Y par rapport à l'erreur-type de y_0	
Variable explicative	Poststrates	Groupes de poststrates	48
$\phi = 0$	1,014	1,045	
$\phi = 0,5$	0,995	1,001	
$\phi = 0,6$	1,000	1,000	
$\phi = 0,7$	1,006	1,001	
$\phi = 0,8$	1,014	1,005	
$\phi = 1,0$	1,046	1,037	
Valeur de Y	2,235		2,294
initiale			

Le tableau 4 présente les estimations de l'EP brutes, Y , et les estimations y_0 du sous-dénombrement net d'un des 48 groupes poststrates. L'estimation de la population totale est la différence entre l'estimation de la population totale dans la poststrate et le chiffre du recensement, divisé par le chiffre du recensement. Nous avons choisi $\phi = 0,6$ comme l'estimateur préféré en fonction des ratios de l'erreur-type brute du tableau 3. Les prédictions et les erreurs-types se ressemblent beaucoup

L'étude a été en partie appuyée par le contrat de coopération 43-3AEU-3-80088 conclu entre l'université de l'Iowa, le National Agricultural Statistics Service et le U.S. Bureau of the Census. Le présent article fait état des résultats de recherche et d'analyse entreprises par les employés du Bureau of the Census et de l'université de l'Iowa. On y fait un examen plus limité que dans les publications officielles du Census Bureau. L'article vise à informer les parties intéressées des résultats de recherche et à encourager la discussion. Nous remercions les lecteurs et les rédacteurs de leurs nombreux commentaires qui ont permis de peaufiner le document.

REMERCIEMENTS

Les estimations des erreurs-types des variables explicatives ont été calculées au moyen de l'approximation de la variance brute de l'annexe B. La moyenne des ratios de l'erreur-type de y_0 à y_0 pour certaines valeurs sélectionnées de ϕ figurent au tableau 3. L'ordonnement des ratios pour les 48 groupes de strates est à peu près le même que pour les 336 poststrates initiales. On constitue un groupe de poststrates en combinant les sept cellules âge-sexe au sein d'un classement donné race-mode d'occupation-degré d'urbanisation-région. En fonction de ces calculs, une valeur ϕ de 0,5 ou de 0,6 est l'estimateur préféré, bien que l'estimation des différences en matière d'efficacité ne soit pas importante. Tout membre de la catégorie ϕ est de loin supérieur à l'estimateur Y initial. L'efficacité de la variance estimée moyenne est d'environ 400% pour les variables explicatives ϕ , relativement aux estimateurs poststrates initiaux.

$$\hat{G}^{\phi} = \hat{\Sigma}_1^{\phi\phi} \hat{\Sigma}^{ww},$$

$$\hat{H}^{\phi} = I - \hat{G}^{\phi} = \hat{\Sigma}_1^{\phi\phi} [\phi \hat{D}^{ec} + (1 - \phi) \hat{\Sigma}^{ec}],$$

$$\hat{\Sigma}^{\phi\phi} = \hat{\Sigma}^{ww} + \phi \hat{D}^{ec} + (1 - \phi) \hat{\Sigma}^{ec},$$

$$\hat{D}^{ec} = \text{diag} \{ \hat{\Sigma}^{ec} \},$$

$$\hat{b}^{\phi} = \left(X' \hat{\Sigma}_1^{\phi\phi} X \right)^{-1} X' \hat{\Sigma}_1^{\phi\phi} Y,$$

et

$$\hat{\Sigma}^{ww} = K_1 \hat{\sigma}_1^2 + K_2 \hat{\sigma}_2^2.$$

La variable explicative de l'équation (31) avec $\phi = 0$ est

l'équation (31) avec $\phi = 1$ est la variable diagonale. Il devrait y avoir une équation $\phi, 0 < \phi < 1$, qui fasse en sorte que la variable explicative ne varie d'une extrême à l'autre.

L'estimation directe de l'EP du nombre total de personnes est la somme pondérée des facteurs de correc-

tion, où les poids sont les chiffres du recensement dans les poststrates. L'erreur-type de l'estimateur direct du total est relativement faible, et l'estimateur direct est considéré comme étant l'estimateur préféré du total. Par conséquent, les variables explicatives du modèle sont établies compte tenu de la contrainte que la somme pondérée des valeurs explicatives est égale à l'estimation directe du total. Par conséquent, la restriction est la suivante

$$\hat{Y}_T = \sum_{336}^t a'_i Y_i = \sum_{336}^t a'_i \tilde{Y}_i,$$

où \hat{Y}_T est l'estimateur direct de l'EP du total, a'_i est la chiffré du recensement dans la même poststrate, et \tilde{Y}_i est la valeur de a'_i a été normalisée pour équivaloir à un. Bateese, Harter et Fuller (1988) ont corrigé les prédictions de manière à ce que les estimateurs satisfassent à la restriction. Chosh et Rao (1994) traitent de telles corrections. Nous employons une méthode qui permet l'estimation directe de la variance des prédictions restreintes.

Nous avons imposé la restriction aux premières variables explicatives selon une méthode qui, approximativement, a établi les meilleures variables explicatives de 335 quantités qui sont estimées comme n'ayant aucune corrélation avec \hat{Y}_T . Disons que $\hat{\Sigma}^{zz}$ est la matrice des covariances estimées de $Y = (Y_1, Y_2, \dots, Y_{336})'$ et définissons

$$CY = (\hat{Y}_T, Y_2 - b_2 \hat{Y}_T, \dots, Y_{336} - b_{336} \hat{Y}_T)',$$

Pour le vecteur des 336 observations, nous avons produit des éléments lissés au moyen de la variable explicative généralisée en (32) pour plusieurs valeurs de ϕ . Soulignons que $\phi = 0$ correspond à la variable explicative de substitution et $\phi = 1$ correspond à la variable explicative diagonale.

4.3 Éléments lissés

variance attribuable à l'estimation de la variance.

considérant H^{ϕ} comme une matrice fixe. Le terme final à droite de l'équation (33) est l'estimateur de la variance définies à l'annexe B. La somme des deux premiers termes où les valeurs de $H^{\phi} = C^{-1} A C H^{\phi}$, et $F_{\beta\beta}^{\phi}$, F_{33}^{ϕ} et F_{44}^{ϕ} sont

$$+ C^{-1} A C [\hat{H}^{\phi} X' X \hat{H}^{\phi} + F_{\beta\beta}^{\phi} + F_{33}^{\phi} + F_{44}^{\phi}] C' A C^{-1}, \quad (33)$$

$$= (I - \hat{H}^{\phi})' \hat{\Sigma}^{ee} (I - \hat{H}^{\phi}) + \hat{H}^{\phi} \hat{\Sigma}^{ww} \hat{H}^{\phi}$$

$$V\{\tilde{y} - y\}$$

L'estimation de la variance de \tilde{y} est la suivante

$$A = \begin{pmatrix} 0 & 0 \\ 0 & I_{335} \end{pmatrix}.$$

où

$$\tilde{y} = Y - C^{-1} A C H^{\phi} (Y - X \hat{b}^{\phi}), \quad (32)$$

explicative de y est la suivante

l'estimateur pour le premier élément de CY , la variable pour les 335 derniers éléments de CY et utilisons \hat{Y}_T comme est CY . Si nous utilisons la variable explicative de modèle de y , alors la variable explicative de modèle de CY .

Si nous disons que \tilde{y} est la variable explicative de ne sont pas en corrélation avec \hat{Y}_T .

I_k est la matrice d'identité $k \times k$, et 0 est un vecteur-colonne ne comprenant que des zéros. Les éléments de CY

$$b^{335} = \begin{pmatrix} 0 \\ I_{335} \end{pmatrix}, \quad \hat{\Sigma}^{aa} (a \hat{\Sigma}^{aa} a)',$$

$$B = \begin{pmatrix} 1 & 0 \\ -b_{335} & I_{335} \end{pmatrix},$$

$$a = (a_1, a_2, \dots, a_{336}),$$

$$T = \begin{pmatrix} 0 \\ a \\ I_{335} \end{pmatrix},$$

$$C = BT,$$

4. APPLICATION AUX DONNÉES DE L'ESTIMATION POSTCENSITAIRE

4.1 Estimation postcensitaire

Le U.S. Bureau of the Census fournit des estimations annuelles de la population des régions en fonction des recensements décennaux et d'autres sources d'information. Afin d'envisager l'utilisation possible des chiffres corrigés du recensement de 1990 dans le cadre du processus d'estimation postcensitaire, le Bureau a examiné les données de l'EP et défini un nouvel ensemble de 357 poststrates.

Les 357 poststrates sont composées de 51 groupes poststrates, dont chacun est sous-divisé en sept catégories d'âge-sexe. Il s'agit des catégories suivantes: (1) les enfants de 0 à 17 ans des deux sexes, (2) les hommes de 18 à 29 ans, (3) les hommes de 30 à 49 ans, (4) les femmes de 18 à 29 ans, (5) les femmes de 30 à 49 ans, (6) les hommes de 50 ans et plus, (7) les femmes de 50 ans et plus. Les facteurs qui définissent les 51 groupes poststrates sont les suivants: race/ethnie (Blancs non hispaniques, Noirs, Hispaniques non Noirs, Amérindiens), le mode d'occupation (propriétaire, locataire), le type de région (zone urbanisée, comptant plus de 250 000 habitants, autre zone urbanisée, zone non urbanisée) et la région (Ouest, Sud, Midwest, Nord-Est). Compte tenu des limites de taille des échantillons, les Amérindiens ont été rassemblés en un groupe poststrate et les Asiatiques ont été divisés en deux groupes poststrates - les propriétaires et les locataires. Des 48 autres groupes poststrates, les 24 premiers groupes reflètent un classement recoupé complet des catégories de Blancs non hispaniques. Les 12 groupes suivants visent les Noirs et comportent un classement recoupé complet du mode d'occupation selon la région pour les zones urbaines comptant plus de 250 000 habitants, mais par ailleurs ne fournissent pas de détails régionaux. Les mêmes 12 groupes poststrates ont servi autant pour les Hispaniques non Noirs que pour les Noirs.

On a obtenu une matrice des covariances de 357 x 357 au moyen de l'algorithme de jaccardité dont on s'est servi pour les 1 392 poststrates de l'EP de 1990. Nous dénotons cette matrice des covariances brute par Σ_{ge} . Dans Hogan (1993), on trouve une description détaillée des 357 poststrates et la motivation pour les établir.

4.2 Modèle de régression

Nous avons éliminé les données sur les Asiatiques et les Amérindiens du processus de lissage. Par conséquent, par groupe minoritaire on entend la combinaison des Noirs et leurs covariances brutes estimatives. L'interaction entre le groupe minoritaire et l'âge-sexe a été incluse dans le modèle de régression après que l'examen des données de 1990 a indiqué que la différence de sous-dénombrement nette entre les Noirs et les non Noirs variait selon le sexe et

le groupe d'âge. Le modèle de régression (1) renferme 21 variables explicatives. Les voici:

1. X_0 = ordonnée à l'origine

2. X_j = variable indicatrice pour les catégories d'âge-sexe: $j = 1, 2, \dots, 6$ dans l'ordre; âges 0-17 ans, hommes 18-29 ans, hommes 30-49 ans, etc. (femmes 50 ans + est la catégorie sans variable)

3. X_7 = variable indicatrice pour les locataires

4. X_8 = variable indicatrice pour les Noirs

5. X_9 = variable indicatrice pour les Hispaniques non Noirs

6. X_j = variable indicatrice pour le genre d'endroit: $j = 10, 11$ pour zone urbanisée de 250 000 habitants+ et autre zone urbaine, respectivement

7. X_j = variable indicatrice pour région: $j = 12, 13, 14$ pour le Nord-Est, le Sud et l'Ouest, respectivement

8. X_j = variable indicatrice pour groupe minoritaire selon l'âge-sexe: $j = 15, \dots, 20$ pour groupe minoritaire de 0-17 ans, hommes minoritaires (18-29 ans), etc.

Les variables X_{12} , X_{13} et X_{14} étaient les proportions du recensement de 1990 des personnes dans le groupe poststrate dans la région particulière pour les groupes poststrates des Noirs et des Hispaniques non Noirs qui ont été combinés d'une région à l'autre.

On a perfectionné le modèle (3) pour l'application empirique. En fonction de l'analyse préliminaire, la structure de l'erreur précisée de w , l'erreur de modèle, est passée de $\Sigma^{ww} = \sigma^2 \mathbf{I}$ à

(30) $\Sigma^{ww} = \mathbf{K}_1 \sigma_1^2 + \mathbf{K}_2 \sigma_2^2$

où \mathbf{K}_1 est une matrice de diagonale $n \times n$ comportant des uns pour les poststrates des groupes minoritaires et des zéros ailleurs, et \mathbf{K}_2 est une matrice de diagonale $n \times n$ comportant des uns pour les poststrates des groupes non minoritaires et des zéros ailleurs. Les variances estimatives sont de $\hat{\sigma}_1^2 = 0,000506$ (0,000140) et de $\hat{\sigma}_2^2 = 0,000112$ (0,000030), où les nombres entre parenthèses sont des erreurs-types. L'erreur-type de la différence est (0,000141). Par conséquent, tout indique que les variances diffèrent pour les deux groupes.

Dans notre discussion, nous avons étudié deux variables explicatives, la variable de substitution en (1) et la variable diagonale en (16). Il est naturel d'envisager une variable de compromis qui aurait la forme suivante

(31)
$$\hat{y}^\phi = X\hat{\beta}^\phi + \hat{G}^\phi(\mathbf{V} - X\hat{\beta}^\phi)$$
$$= \mathbf{V} - \mathbf{H}^\phi(\mathbf{V} - X\hat{\beta}^\phi),$$

où $0 \leq \phi \leq 1$,

$\mathbf{M}^{ee} = \text{diag} \Sigma_{ee}^{zz}, \mathbf{D}^{zz} = \text{diag} \Sigma_{ee}^{zz}$, et la valeur estimée de μ est la suivante

$$\hat{\mu}^d = [\mathbf{J}^d \mathbf{D}^{-1} \mathbf{J}^d - \mathbf{J}^d \mathbf{D}^{-1} \mathbf{J}^d]^{-1} \mathbf{J}^d \mathbf{D}^{-1} \mathbf{y}.$$

Cette variable explicative peut être appelée variable explicative diagonale parce que seuls les éléments de la diagonale de Σ_{ee}^{zz} servent à son établissement.

Les entrées au tableau 1 sont pour $r = 14$. Chaque échantillon est composé d'une sélection aléatoire de \mathbf{w} et d'un échantillon aléatoire de 14 vecteurs \mathbf{e} . On a des

résultats pour les erreurs $w_j \sim \text{NI}(0, 2)$ et pour les erreurs \mathbf{e} . On a des résultats pour les éléments 1 à 4 et pour l'élément 7 sont les ratios pour les éléments de \mathbf{y}^g est définie selon l'équation en (28), par rapport à la variance Monte Carlo de l'élément correspondant de \mathbf{y} pour les erreurs normales. Les ratios pour les éléments de \mathbf{y}^g ne sont pas en corrélation avec les autres éléments. L'élément 7 a une petite variance et l'élément 8 a une grande variance. La perte est grande pour la variable explicative relativement à la moyenne simple pour l'élément 7, et le gain est important pour l'élément 8.

La quatrième colonne du tableau 1 renferme les ratios de la variance de la variable explicative en (29) par rapport à la variance de la moyenne des erreurs normales. Dans tous les cas, la variable explicative de la diagonale est supérieure à la variable explicative générale définie en (28). La différence est relativement constante à environ 30%. La variable de la diagonale n'est pas toujours supérieure à la moyenne simple, mais la perte est faible pour les éléments un, trois et sept. En revanche, les gains relativement à la moyenne simple sont importants pour les éléments six et huit. Les variances Monte Carlo pour les deux variables explicatives sont supérieures aux approximations liées aux équations (15) et (18), sauf pour l'élément 8.

Il est quelque peu surprenant que la méthode de la diagonale ait mieux réussi pour la moyenne simple des erreurs khi-carré que pour celle des erreurs normales. Pour ce qui est de l'erreur khi-carré, les estimations de la moyenne et de la variance sont en corrélation. Par conséquent, en moyenne, les écarts moyens positifs importants sont ramenés vers la moyenne par un montant supérieur à l'écart négatif plus faible. D'après les conjectures établies par le réducteur en chef adjoint, auxquelles nous souscrivons, il s'agit-il d'une des raisons expliquant la performance supérieure de la variable explicative diagonale. En revanche, la méthode de prédiction générale réussit moins bien pour ce qui est de la moyenne simple des erreurs khi-carré que pour celle des erreurs normales. Comme la dernière colonne

du tableau 1 le montre, la méthode de la variable explicative et la diagonale domine uniformément à la fois la moyenne configuration paramétrique avec erreurs khi-carré.

Tableau 1

Ratios de variance Monte Carlo pour variables explicatives régionales alternatives (10 000 échantillons, $r = 14$)

i	Erreurs normales		Erreurs khi-carré	
	$\frac{F(y_j^g - w_j)}{F(y_j^g - w_j)}$	$\frac{F(y_j^g - w_j)}{F(y_j^g - w_j)}$	$\frac{F(y_j^g - w_j)}{F(y_j^g - w_j)}$	$\frac{F(y_j^g - w_j)}{F(y_j^g - w_j)}$
1	0,2414	1,277	1,025	1,430
2	0,3445	1,252	0,875	1,371
3	0,2268	1,351	1,019	1,480
4	0,4771	1,003	0,735	1,099
5	0,4113	0,926	0,876	1,016
6	0,5121	0,913	0,677	0,975
7	0,1449	1,366	1,006	2,261
8	1,1214	0,520	0,384	0,725
				0,371

Les variances Monte Carlo de $\hat{\mu}_0, \hat{\mu}_g$, et $\hat{\mu}_d$ comme estimateurs de μ sont 0,150, 0,273 et 0,146, respectivement. Si les valeurs de Σ_{ee}^{zz} et de σ^2 sont connues, les variances de $\hat{\mu}_0, \hat{\mu}_g$, et $\hat{\mu}_d$ sont 0,149, 0,122 et 0,140, respectivement. Si on utilise une matrice des covariances estimatives pour $\hat{\mu}_g$, on obtient un estimateur ayant une variance supérieure à celle de la moyenne simple.

Dans le cadre du modèle, les variables explicatives sont sans biais quand les erreurs sont normalement distribuées. Les variables explicatives sont biaisées et comprennent des erreurs khi-carré parce que la moyenne de l'échantillon est en corrélation avec la variance empirique. Le tableau 2 donne le biais Monte Carlo divisé par l'erreur-type Monte Carlo de la moyenne. Le biais de la méthode générale est de 20% à 50% supérieur à celui de la méthode de la diagonale. Dans les deux cas, le biais carré ajouté à la variance produit une erreur quadratique moyenne pour ce qui est de la méthode qui est environ 4% à 10% supérieure à la variance. Cette petite étude montre que l'utilisation d'une matrice des covariances estimatives à variabilité importante peut donner lieu à des variables explicatives moins efficaces que la moyenne simple.

Tableau 2

Biais relatif Monte Carlo des variables explicatives régionales alternatives (10 000 échantillons, $r = 14$, erreurs khi-carré)

i	$\frac{\text{Moy.}(\hat{\mu}_g - w_j)}{F(y_j^g - w_j)}$	$\frac{\text{Moy.}(\hat{\mu}_d - w_j)}{F(y_j^g - w_j)}$
1	-0,28	-0,19
2	-0,27	-0,18
3	-0,30	-0,17
4	-0,27	-0,18
5	-0,26	-0,21
6	-0,29	-0,20
7	-0,24	-0,20
8	-0,24	-0,21

configuration donne une gamme de variances d'erreur et une gamme de corrélations entre les estimations. L'estimateur de σ^2 utilisé dans l'étude Monte Carlo est

le suivant

$$\{0, \left[\left[\left\{ \mathbf{A}^{ee} \right\}_{i-1}, \left\{ \mathbf{A}^{(0)} \right\}_{i-1} \right] \times \left[\left\{ \mathbf{A}^{(0)} \right\}_{i-1}, \left\{ \mathbf{A}^{(0)} \right\}_{i-1} \right] \} = \max \{ (k-1)^{-1} \} \quad (22)$$

$$\sum_{l=1}^f (1 - \rho_l) = \sum_{l=1}^f (\underline{\mathbf{A}} - \underline{\mathbf{A}}_l)(\underline{\mathbf{A}} - \underline{\mathbf{A}}_l)' \quad (23)$$

$$\hat{\mu}_0 = k^{-1} \mathbf{J}' \underline{\mathbf{y}}. \quad (24)$$

lié à l'analyse de l'estimateur de la variance.

l'étude Monte Carlo. Les deux vont comme suit

$$(25) \quad (\mathbf{H} - \mathbf{Y})(\mathbf{H} - \mathbf{Y})' = \mathbf{Y}(\mathbf{Y}' - \mathbf{H})$$

survivant

$$(17) \quad \lambda_j = \mu_j + \nu_j, \quad j = 1, 2, \dots, r$$

$$\begin{aligned} \langle \mathbf{Z}^{\alpha\beta} \mathbf{0} \rangle_{\text{pu}} &\sim \int \mathbf{Z}^{\alpha\beta} \mathbf{0} \\ \langle \mathbf{Z}^{\alpha\beta} \mathbf{0} \rangle &\sim \mathbf{M} \end{aligned}$$

$\mathbf{f} = (1, 1, \dots, 1)'$, \mathbf{w} est le vecteur dimensionnel- k des effets

autres sont indépendantes. Le modèle est une

correctes, nous définissons, pour $k = 8$,

$$\begin{bmatrix} e_{18} \\ e_{17} \\ e_{16} \\ e_{15} \\ e_{14} \\ e_{13} \\ e_{12} \\ e_{11} \\ e_{10} \\ e_9 \\ e_8 \end{bmatrix} = \begin{bmatrix} 1.0n_{17} & 1.6n_{16} & 0.9n_{15} & 1.0n_{14} & 1.5n_{13} & 0.4n_{12} & 0.9n_{11} & 0.6n_{10} & 1.6n_9 & 1.0n_8 & 2.8n_7 \end{bmatrix}$$

ou w_i sont des variables aléatoires indépendantes. Les w_i , $i = 1, 2, \dots, 8$, sont des variables aléatoires $NI(0, 0.36)$, où $NI(\mu, \sigma^2)$ dénote des variables aléatoires indépendantes normales dont la moyenne est μ et la variance σ^2 . Cette

$$(92) \quad \left(\int_{\mathcal{Z}} \mathbf{H} - \mathbf{A} \right)^{\mathcal{S}} \mathbf{H} - \mathbf{A} = \mathcal{L}$$

$$\mathbf{J}^g \mathbf{J}^{\bar{g}} \mathbf{J}^{\bar{g}} \mathbf{J}^g = \mathbf{J}^g \mathbf{J}^{\bar{g}} \mathbf{J}^g \mathbf{J}^{\bar{g}} \quad (27)$$

et μ_g est l'estimateur estimatif des moindres carrés généralisés de μ .
La deuxième variable explicative est la suivante

$$(67) \quad \left(\int_{\Gamma} \left(\mathbf{J}^p \mathbf{H} - \mathbf{A} \right)^p \right) \mathbf{H} - \mathbf{A} = \mathbf{A}$$

Puis, on estime σ^2 selon le modèle en (1), (2) et (5) en considérant l'estimateur de Σ^{ee} comme la vraie valeur de Σ .

La variable explicative de substitution peut être

$$\hat{y} = X\beta + \sigma^2 \Sigma^{-1}(\mathbf{Y} - X\beta), \quad (11)$$

où

$$\hat{\beta} = (X' \hat{\Sigma}^{-1} X)^{-1} X' \hat{\Sigma}^{-1} \mathbf{Y} \quad (12)$$

est l'estimateur des moindres carrés généralisés estimés de β ,

$$\hat{\Sigma}^{zz} = \mathbf{I}\sigma^2 + \hat{\Sigma}^{ee} \quad (13)$$

$\hat{\Sigma}^{ee}$ est un estimateur de Σ^{ee} , et $\hat{\sigma}^2$ est un estimateur de σ^2 .

L'estimateur de σ^2 peut se fonder sur des méthodes de vraisemblance ou d'analyse de la variance. Ne retenant que les termes de l'extension de l'erreur de Taylor en (11) qui sont les erreurs dans les estimateurs de base, nous avons l'équation suivante

$$\hat{y} - y = \mathbf{e} - \mathbf{H}'\mathbf{z} + \mathbf{H}'\mathbf{X}(\hat{\beta} - \beta) + (\hat{\sigma}^2 - \sigma^2) \mathbf{H}'\Sigma^{-1}\mathbf{z} - \mathbf{G}'(\hat{\Sigma}^{ee} - \Sigma^{ee})\Sigma^{-1}\mathbf{z}, \quad (14)$$

où $\mathbf{H}' = \Sigma^{ee}\Sigma^{-1}$ et $\mathbf{G}' = \mathbf{I} - \mathbf{H}' = \sigma^2 \Sigma^{-1}$. Si on suppose que la valeur de Σ^{ee} est distribuée comme un multiple d'une matrice Wishart ayant des degrés de liberté d_e , si on ne tient pas compte de la covariance entre $\hat{\sigma}^2$ et $\hat{\Sigma}^{ee}$, si on calcule les espérances comme si les valeurs de $\hat{\sigma}^2$ et de \mathbf{z} sont indépendantes et si on calcule les espérances comme si les valeurs de \mathbf{z} et de $\hat{\Sigma}^{ee}$ sont indépendantes, on obtient, à partir de l'équation en (14), comme approximation de la variance de $\hat{y} - y$ l'équation suivante

$$\mathbf{V}\{\hat{y} - y\} = \Sigma^{ee}\mathbf{G} + \mathbf{H}'\mathbf{X}\mathbf{V}\beta\mathbf{H} + \mathbf{I}\sigma^2 + \mathbf{I}\sigma^4, \quad (15)$$

où

$$\mathbf{V}\beta = \mathbf{V}\{\hat{\beta}\} = (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1} + d_e^{-1}\mathbf{u}'\{\Sigma^{-1}\Sigma^{ee}\Sigma^{-1}\mathbf{u}\} \mathbf{I}\Sigma^{ee}\mathbf{I},$$

$$\mathbf{I}\sigma^4 = \mathbf{H}'\Sigma^{-1}\mathbf{H}\mathbf{I}\sigma^4,$$

$$\mathbf{I}\sigma^4 = d_e^{-1}\sigma^4 \Sigma^{-1}\Sigma^{ee}\Sigma^{-1}\mathbf{u}'\{\Sigma^{-1}\Sigma^{ee}\Sigma^{-1}\mathbf{u}\},$$

$$\mathbf{I} = (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1}$$

et $\mathbf{I}\sigma^4 = \mathbf{V}\{\hat{\sigma}^2\}$ est la variance de $\hat{\sigma}^2$. Le terme $\Sigma^{ee}\mathbf{G}$ est la matrice des covariances de prédiction si tous les paramètres sont connus. Les autres trois termes de (15) sont les apports à la variance attribuables à l'estimation de β , σ^2 , et Σ^{ee} , respectivement. Le deuxième terme dans l'expression $\mathbf{V}\{\hat{\beta}\}$ est une estimation brute de l'augmentation de la variance de $\hat{\beta}$ attribuable à l'utilisation d'un estimateur de Σ^{zz} plutôt que de Σ^{zz} pour établir $\hat{\beta}$.

Si la dimension de Σ^{zz} est grande et que les degrés de liberté, d_e , ne sont que légèrement supérieurs à la dimension, alors la deuxième partie de la variance de $\hat{\beta}$ et le terme $\mathbf{I}\sigma^4$ peuvent contribuer considérablement à la variance. Cela est particulièrement vrai si la valeur de σ^2 est relativement faible par rapport aux éléments de la diagonale de Σ^{ee} . L'étude Monte Carlo à la section suivante montre que l'apport à la variance calculé approximativement par ces termes peut être important.

Une prédiction qui réduit l'effet de l'erreur d'estimation dans Σ^{ee} n'utilise que les éléments de la diagonale de Σ^{ee} dans la composante de réduction. Posons l'équation suivante

$$\hat{y}^p = X\beta^p + \sigma^2 \mathbf{D}^{-1}(\mathbf{Y} - X\beta^p), \quad (16)$$

où

$$\beta^p = (\mathbf{X}'\mathbf{D}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}^{-1}\mathbf{Y},$$

$$\mathbf{D}^{zz} = \text{diag}(\Sigma^{ee} + \mathbf{I}\sigma^2),$$

$\hat{\sigma}^2$ est un estimateur de σ^2 et $\text{diag}(\mathbf{A})$ est la matrice diagonale composée des éléments de la diagonale de \mathbf{A} . En ne retenant que les termes principaux de l'extension de l'erreur de Taylor en (16), on obtient l'équation suivante

$$\hat{y}^p - y = -(\mathbf{w} - \mathbf{G}'\mathbf{z}) + \mathbf{H}'\mathbf{X}(\hat{\beta} - \beta) + (\hat{\sigma}^2 - \sigma^2) \mathbf{H}'\mathbf{D}^{-1}\mathbf{z} - \mathbf{G}'(\hat{\mathbf{D}}^{ee} - \mathbf{D}^{ee})\mathbf{D}^{-1}\mathbf{z}, \quad (17)$$

où $\mathbf{D}^{zz} = \text{diag}(\Sigma^{zz})$, $\mathbf{G}^p = \mathbf{D}^{-1}\sigma^2$, $\mathbf{H}^p = \mathbf{I} - \mathbf{G}^p$, et $\mathbf{D}^{ee} = \text{diag}(\Sigma^{ee})$. Si les valeurs de \mathbf{w} et de \mathbf{e} sont normalement distribuées, et si $\hat{\sigma}^2$ et $\hat{\mathbf{D}}^{zz}$ sont des estimateurs quadratiques, alors les valeurs de $\hat{\sigma}^2$ et de $\hat{\mathbf{D}}^{zz}$ ne sont pas en corrélation avec \mathbf{z} . Le x -ième élément de $\mathbf{w} - \sigma^2 \mathbf{D}^{-1}\mathbf{z}$ n'est pas en corrélation avec le x -ième élément de \mathbf{z} , sans nécessiter ne pas être en corrélation avec le vecteur \mathbf{z} . Si on ne tient pas compte de cette corrélation possible, si on suppose que Σ^{ee} est une matrice Wishart ayant des degrés de liberté d_e , et si on ne tient pas compte de la covariance entre $\hat{\sigma}^2$ et $\hat{\Sigma}^{ee}$, on obtient une approximation de la variance de $\hat{y}^p - y$ à partir de l'équation en (17) par l'équation suivante

$$\mathbf{V}\{\hat{y}^p - y\} = \mathbf{H}'\mathbf{H}^p\mathbf{G}^p + \mathbf{G}^p\sigma^2 + \mathbf{H}'\mathbf{X}\mathbf{V}\beta\mathbf{H} + \mathbf{I}\sigma^4 + \mathbf{I}\sigma^4,$$

(18)

$$\mathbf{G}^p = \mathbf{D}^{-1}\sigma^2, \quad \mathbf{H}^p = \mathbf{I} - \mathbf{G}^p,$$

$$\mathbf{V}\beta = (\mathbf{X}'\mathbf{D}^{-1}\mathbf{X})^{-1} + d_e^{-1}\mathbf{u}'\{\Sigma^{-1}\Sigma^{ee}\Sigma^{-1}\mathbf{u}\},$$

(19)

$$\mathbf{I}\sigma^4 = \mathbf{H}'\mathbf{D}^{-1}\Sigma^{zz}\mathbf{D}^{-1}\mathbf{H}\mathbf{I}\sigma^4,$$

(20)

$$\mathbf{I}\sigma^4 = d_e^{-1}\sigma^4 \mathbf{G}'\mathbf{G},$$

à un estimateur instable de la matrice des covariances d'erreur. Bien qu'on ait lissé l'estimation des variances d'erreur, on juge que les variances estimatives des combinaisons linéaires peuvent toujours être importantes. Il estime que les variances estimatives sont grandes parce que les estimations directes pour de nombreux blocs sont de zéro. Le Secretary of Commerce a fini par décider d'utiliser les chiffres non corrigés dans le recensement décennal. On a décidé d'étudier ultérieurement la possibilité d'utiliser des chiffres corrigés à d'autres fins, comme pour le programme d'estimations postcensitaires du Bureau.

Nous examinons d'autres estimateurs de lissage pour les facteurs de correction, en insistant sur l'effet de l'estimation de la matrice des covariances du vecteur des facteurs de correction estimatifs. Dans la partie empirique de notre étude, nous établissons des estimations à partir des données du recensement de 1990.

2. MODÈLE DE LISSAGE

Le modèle retenu pour l'établissement des variables explicatives est le modèle des composantes de variance multidimensionnelles. Un certain nombre d'auteurs ont étudié des modèles étroitement liés qui permettent d'estimer le lissage d'un ensemble de valeurs inconnues. Fay et Herriot (1979) ont suggéré d'utiliser le modèle dans le cadre d'une méthode d'estimation régionale. Battese, Harter et Fuller (1988) ont appliqué le modèle des composantes de variance à l'estimation de la zone cultivée. Ericksen et Kadane (1985), Cressie (1992), de même qu'Ericksen, Kadane et Tukey (1989) ont suggéré des méthodes de lissage pour corriger le recensement. Singh, Gambino et Mantel (1994) abordent une gamme de méthodes s'appliquant aux régions. On trouve dans Efron et Morris (1972), de même que dans Morris (1983) de bonnes discussions sur certains aspects théoriques de base. Kackar et Harville (1984), Peixoto et Harville (1986), Fay (1987), Fuller et Harter (1987), Hulting et Harville (1991), Ghosh (1992), de même que Prasad et Rao (1990) traitent de l'estimation de la variance quant à de telles méthodes. Le document de Ghosh et Rao (1994) est un exposé de synthèse.

Dans le modèle des composantes de variance multidimensionnelles, le vecteur des valeurs vraies à prédire est le suivant

$$(1) \quad y = X\beta + w,$$

où y est un vecteur-colonne dimensionnel- n , X est une matrice $n \times k$ de caractéristiques observables, w est un vecteur-colonne dimensionnel- n des effets aléatoires et β est un vecteur-colonne d'inconnues dimensionnel- k . Le vecteur

$$(2) \quad Y = y + e,$$

Y est observé, si

où $G^{zz} = \Sigma^{-1} \sigma^2$ et $\Sigma^{zz} = I\sigma^2 + \Sigma^{ee}$ est la matrice des covariances $n \times n$ de $z = w + e$. Selon le modèle de distribution normale défini en (1), (2) et (5) et compte tenu des paramètres σ^2 , Σ^{ee} , β connus, la variable explicative du côté droit de l'équation (6).

De façon générale, certains des paramètres sont inconnus. Prenons d'abord le cas où la valeur de β est inconnue. Disons que β est un estimateur de β , où

$$(6) \quad E\{y | X\} = X\beta + G^{zz}(Y - X\beta),$$

et M est une matrice $n \times n$. Si la valeur de M est établie

$$(7) \quad \hat{\beta} = (X'M^{-1}X)^{-1}X'M^{-1}Y,$$

où $K = (I - G')X(X'M^{-1}X)^{-1}X'M^{-1} + G'$. Par conséquent, si les valeurs de M et de G sont établies,

$$(8) \quad V\{y - y\} = (K - I)(K - I)'\sigma^2 + K\Sigma^{ee}K',$$

Si le modèle en (1), (2) et (3) tient, et si les valeurs de Σ^{ee} et de σ^2 sont connues, alors en remplaçant B par

$$(10) \quad G^{zz} = \Sigma^{-1} \sigma^2$$

en (4), on obtient la meilleure variable explicative linéaire sans biais de y . Voir Henderson (1950), Harville (1976) et Robinson (1991). Si Σ^{ee} et σ^2 sont également inconnues, il est naturel d'utiliser les estimateurs de Σ^{ee} et de σ^2 pour établir une estimation de la meilleure variable explicative linéaire sans biais. Très souvent, l'estimateur de Σ^{ee} est associé à la méthode ayant servi à établir l'estimateur de Y .

Estimation des facteurs de correction au recensement

C.T. ISAKI, J.H. TSAY et W.A. FULLER¹

RÉSUMÉ

À partir d'une méthode des composantes de variance et d'une structure estimative des erreurs de covariance, on a établi les variables explicatives des facteurs de correction pour le recensement décennal de 1990. On soupçonne que la variabilité des covariances estimatives explique certaines anomalies dans l'estimation de régression et les facteurs de correction estimés. Nous avons étudié des méthodes de prédiction alternatives et proposé une façon de faire qui est moins sensible à la variabilité de la matrice des covariances estimatives. La méthode proposée est appliquée à un ensemble de données composé de 336 facteurs de correction à partir de l'enquête postcensitaire de 1990.

MOTS CLÉS: Composantes de variance; estimation régionale; sous-dénombrement; recensement décennal; lissage.

1. INTRODUCTION

Bien que l'objectif du recensement de la population soit de consigner des données pour toutes les personnes, on sait bien depuis longtemps qu'en pratique il n'en est rien. D'après les études postcensitaires liées au recensement américain de 1970 et de 1980, le taux de couverture n'est pas le même pour différents groupes démographiques. Voir le U.S. Bureau of the Census (1988).

En 1990, on s'est servi d'une enquête postcensitaire (BP), fondée sur une estimation à système double (ou saisie-resaisie), pour produire des estimations à l'égard de 1392 sous-divisions de la population totale des États-Unis au moment du Recensement de 1990. L'échantillon de l'BP comprenait environ 377 000 personnes dans environ 5 200 blocs d'échantillon. Les personnes échantillonnées étaient divisées en poststrates définies selon les divisions géographiques du pays, le mode d'occupation, la taille de l'endroit, la race, le sexe et l'âge dans le cadre desquelles les catégories du mode d'occupation sont les propriétaires et les locataires de résidences, et la taille de l'endroit constitue une mesure de l'urbanisation. Les sous-divisions sont appelées des poststrates. Le ratio de l'estimation de l'BP par rapport au total du recensement, appelé le facteur de correction supérieur à un correspond à un sous-dénombrement estimatif et un élément inférieur à un correspond à un surdénombrement estimatif.

Parce qu'on prévoyait des variances d'échantillonnage relativement importantes pour les ratios individuels, une technique de lissage fondée sur les composantes de variance et un modèle de régression ont permis de créer les estimations finales des facteurs de correction. Les éléments de la matrice des covariances d'erreur dont on s'est servi dans le modèle prédictif ont été estimés au moyen d'un algorithme de jackknife (Fay 1990).

Dans le modèle de régression, on a choisi les variables explicatives selon un algorithme de sélection du meilleur

sous-ensemble. Certaines variables explicatives ont été imposées au modèle. Par exemple, dans la région du Midwest, les dix variables explicatives imposées au modèle étaient les Noirs, les Hispaniques, les locataires, le groupe des 0-9 ans, le groupe des 10-19 ans, le groupe des 20-29 ans, le groupe des 30-44 ans, le groupe des 45-64, les hommes âgés de 10-19 ans et les hommes âgés de 20-64 ans. La plupart des variables étaient des variables indicatrices, mais certaines étaient des proportions. Par exemple, on recourait à une variable «pourcentage de Noirs» quand les Noirs et les Hispaniques étaient regroupés dans une seule poststrate. On a sélectionné neuf autres variables à inclure dans le modèle à partir d'un algorithme de régression du meilleur sous-ensemble. Les variables comprenaient le taux de questionnaires renvoyés par la poste, le taux de substitution, le genre d'endroit et six variables d'interaction de la race selon l'âge et de la race selon le mode d'occupation. Le taux de questionnaires renvoyés par la poste constitue la fraction renvoyée des questionnaires du recensement distribués par la poste; le taux de substitution est la fraction des ménages visés par le recensement qui a été entièrement remplacé par les ménages répondants.

On a appliqué la technique de lissage aux ratios poststrates selon les régions du pays. Les facteurs de correction devaient s'appliquer aux chiffres du recensement dans les poststrates appropriées pour que les estimations de population tiennent compte du sous-dénombrement et du surdénombrement. Dans Hogan (1992), on trouve un aperçu de l'BP. Isaki et coll. (1991) fournissent une description détaillée des résultats du lissage des ratios poststrates.

Fay (1992), dans un document portant sur les facteurs de correction établis à partir de l'BP de 1990, a relevé certains résultats perturbateurs. Il souligne que certains des coefficients de régression estimatifs dans le modèle diffèrent considérablement selon la forme de la matrice des covariances estimatives ayant servi à établir l'estimateur des moindres carrés généralisés estimés. Fay conjecture que les grandes différences entre les coefficients peuvent être attribuables

¹ C.T. Isaki et J.H. Tsay, Statistical Research Division, Bureau of the Census, Washington, D.C. 20233, U.S.A.; W.A. Fuller, Department of Statistics, Iowa State University, Ames, IA 50010, U.S.A.

la statistique des outils permettant de résoudre des problèmes pratiques a été le trait marquant du recensement des États-Unis. Les problèmes pratiques que doit résoudre le gouvernement concernent sûrement parmi les plus fondamentaux et Mahalanobis réservait des ressources statistiques pour la résolution de ce genre de problème. Le U.S. Census Bureau a lui aussi depuis longtemps la réputation d'offrir des solutions pratiques et rentables pour résoudre les problèmes épineux que pose le recensement.

BIBLIOGRAPHIE

- BAILLAR, B.A. (1969). Evaluation and Research Program of the U.S. Censuses of Population and Housing, 1960: The Effect of Interviewers and Crew Leaders. Series ER 60 No. 7. Washington, DC: U.S. Bureau of the Census.
- EDMONSTON, B., et SCHULTZE, C.V. (1993). *Modernizing the U.S. Census*. Washington DC: National Academy Press, 34-35.
- HANSON, R.H., et MARKS, E.S. (1958). The influence of the interviewer on the accuracy of survey results. *Journal of the American Statistical Association*, 53, 635-655.
- MAHALANOBIS, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, 325-378.
- MAHALANOBIS, P.C. (1950). Why Statistics? *Sankhya*, 10, 195-228.
- MAHALANOBIS, P.C., et LAHIRI, D.B. (1961). Analysis of errors in censuses and surveys with special reference to experience in India. *Bulletin de l'Institut Internationale de Statistique*, 38, 2, 401-433 (republié dans *Sankhya*, 23, 325-358).
- MICHAEL, R.T., GAGNON, J.H., LAUMANN, E.O., et KOLATA, G. (1994). *Sex in America, A Definitive Survey*. New York: Little Brown and Co.
- RUDRA, A. (1996). *Prasanna Chandra Mahalanobis: A Biography*. New York: Oxford University Press.
- WAKSBERG, J., et HANSON, R. (1965). Sampling Applications in Censuses of Population and Housing. U.S. Bureau of the Census, Technical Paper No. 13.

Pour les secteurs de 50 000 habitants, qui existe dans la plupart des grandes villes, les résultats montrent l'effet dévastateur dû à la non-corrrection pour le sous-dénombrement dans le cas des minorités nombreuses. Pour les totaux importants, la variance d'échantillonnage n'a pratiquement aucun effet sur la valeur de la REQ_M, mais dans tous les cas, l'amélioration résultant de la correction, avec ou sans échantillonnage, est égale ou supérieure à 83%. L'erreur supplémentaire due à l'échantillonnage est négligeable.

Malheureusement, dans nombre de collectivités minoritaires, on observe à la fois un faible taux de retour des questionnaires par la poste et un sous-dénombrement. Le taux de retour des questionnaires par la poste est égal ou inférieur à 50% et il se pourrait que l'on doive augmenter la

Durant la deuxième moitié du siècle, le Census Bureau a pris l'habitude d'appliquer des méthodes statistiques, dans la mesure du possible, pour augmenter l'exactitude des données du recensement décennal et en réduire le coût. En se servant des méthodes utilisées traditionnellement par le Census Bureau, à savoir un modèle de l'erreur quadratique moyenne, on constate que la correction améliore la qualité des totaux de recensement, même dans le cas des petits territoires de recensement, si l'on suppose qu'il existe une variance même minime de réponse. On constate aussi qu'il y a lieu d'être prudent si l'on veut recourir à l'échantillonnage pour faire le suivi. Il pourrait être contre-indiqué d'appliquer cette méthode aux territoires dont la population est égale ou inférieure à 2 500 habitants, au même titre qu'il faut éviter l'échantillonnage pour évaluer certaines caractéristiques démographiques dans ces petits territoires.

Étant donné la controverse actuelle au sujet du recensement, il est bon de se souvenir de l'attitude de Mahalanobis. Son utilisation ingénieuse des sous-échantillons superposés nous a certes donné la capacité d'estimer la variance de réponse dans le cadre du recensement, mais son insistance à voir dans l'échantillonnage et

6. CONCLUSION

les territoires où les caractéristiques sont fortement corrélées au sous-dénombrement.

Pour les territoires de 10 000 habitants, nous constatons que la correction produit une amélioration manifeste du biais, mais que les chiffres corrigés par échantillonnage restent nettement plus élevés que les chiffres corrigés sans échantillonnage. Cependant, en cas de chiffres non corrigés, l'échantillonnage augmente la REQM, mais les nombres non corrigés diffèrent peu. On enregistre une augmentation de 15% de la REQM quand la moitié seulement de la population retourne le questionnaire de recensement et une augmentation de 9% quand la proportion est de 70%.

Enfin, si nous examinons le territoire de 50 000 habitants, nous constatons que le biais le terme dominant de REQM

Tableau 5
REQM prévue des fréquences estimatives par cellule pour les estimations de population des collectivités afro-américaines, amérindiennes et hispaniques fondées sur les chiffres non corrigés de recensement, les chiffres corrigés de recensement et un échantillon au 1/4 de ménages n'ayant pas retourné le questionnaire par la poste

A. Territoire de 2 500 habitants, calcul de la REQM fondé sur									
Fréquence par cellule		Non corrigés		Corrigés		Non corrigés		Corrigés	
Pas d'échantillonnage des ménages qui ne retournent pas le questionnaire par la poste		Echantillon au 1/4 et taux de retour par la poste de 0,50		Echantillon au 1/4 et taux de retour par la poste de 0,70		Pas d'échantillonnage des ménages qui ne retournent pas le questionnaire par la poste		Echantillon au 1/4 et taux de retour par la poste de 0,70	
15	3	3	6	6	5	15	3	3	5
50	6	6	11	11	9	50	6	9	13
100	9	9	15	15	26	100	9	13	26
500	26	26	36	30	29	500	26	33	29
750	31	20	46	34	31	750	31	42	31
1 000	45	21	54	37	31	1 000	45	51	31
1 500	64	21	70	37	31	1 500	64	68	31
B. Territoire de 10 000 habitants, calcul de la REQM fondé sur									
Fréquence par cellule		Non corrigés		Corrigés		Non corrigés		Corrigés	
Pas d'échantillonnage des ménages qui ne retournent pas le questionnaire par la poste		Echantillon au 1/4 et taux de retour par la poste de 0,50		Echantillon au 1/4 et taux de retour par la poste de 0,70		Pas d'échantillonnage des ménages qui ne retournent pas le questionnaire par la poste		Echantillon au 1/4 et taux de retour par la poste de 0,70	
50	6	6	11	11	9	50	6	9	13
100	9	9	15	15	26	100	9	13	26
500	26	26	36	30	29	500	26	33	29
750	31	20	46	34	31	750	31	42	31
1 000	45	21	54	37	31	1 000	45	51	31
1 500	64	21	70	37	31	1 500	64	68	31
C. Territoire de 50 000 habitants, calcul de la REQM fondé sur									
Fréquence par cellule		Non corrigés		Corrigés		Non corrigés		Corrigés	
Pas d'échantillonnage des ménages qui ne retournent pas le questionnaire par la poste		Echantillon au 1/4 et taux de retour par la poste de 0,50		Echantillon au 1/4 et taux de retour par la poste de 0,70		Pas d'échantillonnage des ménages qui ne retournent pas le questionnaire par la poste		Echantillon au 1/4 et taux de retour par la poste de 0,70	
50	6	6	11	11	9	50	6	9	13
100	9	9	15	15	28	100	9	13	28
500	21	19	38	33	39	500	21	34	39
1 000	48	26	60	45	51	1 000	48	56	51
2 000	87	35	100	60	64	2 000	87	95	64
5 000	205	43	214	75	64	5 000	205	210	64
REQM prévue des fréquences estimatives par cellule pour les estimations de population des collectivités afro-américaines, amérindiennes et hispaniques fondées sur les chiffres non corrigés de recensement, les chiffres corrigés de recensement et un échantillon au 1/4 de ménages n'ayant pas retourné le questionnaire par la poste									
Fréquence par cellule		Non corrigés		Corrigés		Non corrigés		Corrigés	
Pas d'échantillonnage des ménages qui ne retournent pas le questionnaire par la poste		Echantillon au 1/4 et taux de retour par la poste de 0,50		Echantillon au 1/4 et taux de retour par la poste de 0,70		Pas d'échantillonnage des ménages qui ne retournent pas le questionnaire par la poste		Echantillon au 1/4 et taux de retour par la poste de 0,70	
250	17	14	26	23	21	250	17	23	21
500	28	19	39	33	29	500	28	35	29
1 000	48	27	62	47	40	1 000	48	57	40
2 500	109	42	124	73	63	2 500	109	118	63
5 000	208	58	224	101	86	5 000	208	218	86
10 000	407	77	422	134	115	10 000	407	416	115
25 000	1 004	97	1 014	168	144	25 000	1 004	1 010	144

pour tous les effectifs de cellule sauf les plus faibles. Si le total que nous essayons d'estimer est égal ou supérieur à 5 000, l'échantillonnage augmente la valeur de la REQM, mais la correction des chiffres, conjuguée à l'échantillonnage, donne de meilleurs résultats que les chiffres non corrigés sans échantillonnage.

Le tableau 5 est comparable, mais axé sur une population principalement minoritaire. Comme au tableau 3, le biais relatif est de 4%, reflétant le taux moyen de sous-dénombrement observé pour les minorités. Dans ce tableau, les ET calculées pour les totaux non corrigés sont beaucoup plus comparables, même pour les petits territoires, à cause de l'effet important du terme de biais sur la REQM.

Tableau 4

REQM prévue des fréquences estimatives par cellule pour les estimations de population fondées sur les chiffres corrigés et non corrigés de recensement et sur un échantillon au 1/4 de ménages qui ne retournent pas le questionnaire par la poste

A. Territoire de 2 500 habitants, calcul de la REQM fondé sur									
Fréquence par cellule		Non corrigés		Corrigés		Non corrigés		Corrigés	
Pas d'échantillonnage des ménages qui ne retournent pas le questionnaire par la poste		Echantillon au 1/4 et taux de retour par la poste de 0,50		Echantillon au 1/4 et taux de retour par la poste de 0,70					
15	3	3	6	6	11	6	5	5	5
50	6	6	11	11	11	11	9	9	9
100	9	9	15	15	15	15	13	13	13
500	20	20	17	32	30	28	26	26	26
750	25	25	20	38	34	33	29	29	31
1 000	29	21	42	47	37	37	37	31	31
1 500	37	21	47	37	37	43	31		
B. Territoire de 10 000 habitants, calcul de la REQM fondé sur									
Fréquence par cellule		Non corrigés		Corrigés		Non corrigés		Corrigés	
Pas d'échantillonnage des ménages qui ne retournent pas le questionnaire par la poste		Echantillon au 1/4 et taux de retour par la poste de 0,50		Echantillon au 1/4 et taux de retour par la poste de 0,70					
15	3	3	6	6	11	6	5	5	5
50	6	6	11	11	11	11	9	9	9
100	9	9	15	15	15	13	13	13	13
500	20	20	17	32	30	28	26	26	26
750	25	25	20	38	34	33	29	29	31
1 000	29	21	42	47	37	37	37	31	31
1 500	37	21	47	37	37	43	31		
C. Territoire de 50 000 habitants, calcul de la REQM fondé sur									
Fréquence par cellule		Non corrigés		Corrigés		Non corrigés		Corrigés	
Pas d'échantillonnage des ménages qui ne retournent pas le questionnaire par la poste		Echantillon au 1/4 et taux de retour par la poste de 0,50		Echantillon au 1/4 et taux de retour par la poste de 0,70					
250	15	14	24	24	24	21	20	20	20
500	22	19	35	35	33	30	29	29	29
1 000	34	27	51	47	47	45	40	40	40
2 500	65	42	89	73	73	80	63	63	63
5 000	116	58	142	101	132	132	86	86	86
10 000	214	77	241	134	231	231	115	115	115
25 000	509	97	527	168	520	520	144	144	144

Tableau 3
REQM prévue des fréquences estimatives par cellule pour les estimations de population des collectivités afro-américaines, amérindiennes et hispaniques d'après les chiffres de recensement corrigés et non corrigés pour le sous-dénombrement

Territoire de 2 500 habitants, REQM fondée sur			Territoire de 10 000 habitants, REQM fondée sur			Territoire de 50 000 habitants, REQM fondée sur		
Fréquence	Non corrigés	Corrigés	Fréquence	Non corrigés	Corrigés	Fréquence	Non corrigés	Corrigés
par cellule			par cellule			par cellule		
15	3	3	50	6	6	250	17	14
50	6	6	100	9	9	500	28	19
100	9	8	200	15	12	1 000	48	27
500	26	17	500	21	19	2 500	109	42
750	31	20	1 000	48	26	5 000	208	58
1 000	45	21	2 000	87	35	10 000	407	77
1 500	64	21	5 000	205	43	25 000	1004	97

Nota: Les valeurs sont calculées en supposant que le biais relatif de réponse est égal à 4% si le chiffre de recensement n'est pas corrigé et nul s'il l'est. Dans le cas des chiffres corrigés ainsi que non corrigés de recensement, la variance de réponse est égale au quart de ce que serait la variance d'échantillonnage observée pour un échantillon au 1/4.

On a supposé, pour établir le tableau 2, que le biais relatif de réponse était de 2%, d'après l'estimation globale du sous-dénombrement de 1,6% obtenue, pour le Recensement de 1990. Cependant, puisque le sous-dénombrement est plus important pour les populations minoritaires que pour les autres, comparons des chiffres corrigés et non corrigés de recensement pour lesquels le biais relatif est de 4%. (Les estimations du sous-dénombrement au Recensement de 1990 se chiffrent à 4,4% pour les Afro-américains, à 4,5% pour les Amérindiens, à 5,0% pour les Hispaniques et à 2,3% pour les Asiatiques.)

Le tableau 3 montre que la REQM calculée pour les collectivités minoritaires pour des territoires de recensement de 2 500, 10 000 et 50 000 habitants. Bien que la REQM calculée pour les chiffres corrigés de recensement ne varie pas, puisque le biais a été éliminé, la REQM obtenue pour les chiffres non corrigés est nettement plus grande. Comme il fallait s'y attendre, le gain d'exactitude lié à la correction est beaucoup plus important dans le cas des collectivités minoritaires. Par exemple, comme on l'a montré plus haut, l'erreur qui entache la détermination du nombre d'hommes dans une collectivité non minoritaire de 10 000 personnes serait de l'ordre de 109 en cas de non-correction et de 43 en cas de correction. Pour une communauté minoritaire, les erreurs seraient de 205 et de 43, respectivement. Pour un territoire de recensement plus grand, disons de 50 000 habitants, l'amélioration de l'exactitude est spectaculaire, même pour une petite cellule dont l'effectif est égal à 1 000.

Maintenant, supposons que nous abrogeons la loi de 1976 qui stipule que l'on ne peut recourir à l'échantillonnage pour établir les chiffres sur lesquels se fonde la répartition des ressources. Imaginons un recensement durant lequel, à une date particulière, on échantillonne les logements pour lesquels un questionnaire de recensement n'a pas été retourné.

Ce modèle tient compte de deux composantes de la variance, soit la variance de réponse et la variance

d'échantillonnage. La variance d'échantillonnage est basée uniquement sur l'univers de la non-réponse. Soit R , le taux de non-réponse et M , la population de ménages non répondants. Alors, $M = RV$. Le total pour lequel nous essayons d'estimer la variance d'échantillonnage est $S = PM$. La relation entre S , c'est-à-dire la partie échantillonnée du total et T , c'est-à-dire le total, passe par R . $S = PM = P(RN) = RT$.

Donc, la variance d'échantillonnage est égale à $3MPQ = 3PQRN$, si l'on suppose que l'on sélectionne un échantillon au 1/4 de non-répondants. Cette fraction d'échantillonnage pourrait facilement être remplacée par une fraction plus grande, mais elle suffit pour les besoins de l'exemple.

Au tableau 4, trois éléments contribuent à la REQM. Deux d'entre eux sont les termes que nous avons considérés lors de la description antérieure où l'on n'envisageait pas l'échantillonnage des questionnaires non retournés par la poste. Maintenant, nous avons à faire un troisième terme, qui exprime la variance d'échantillonnage associée à l'échantillon de questionnaires non retournés par la poste. En cas de correction, seul le terme de biais devient nul, l'autre diminue à mesure que l'effectif de la cellule augmente, mais ils ne disparaissent pas.

Le tableau 4 montre les REQM, dans le cas d'un recensement sans échantillonnage des ménages qui ne retournent pas le questionnaire, avec ou sans correction pour le sous-dénombrement pour un échantillon au 1/4 de ménages n'ayant pas retourné le questionnaire par la poste, dans le cas où la moitié seulement de la population les retourne par la poste et dans celui où 70% les renvoie par la poste pour les tailles de territoire de recensement examinées, c'est-à-dire 2 500, 10 000 et 50 000 habitants. Le cas de l'absence d'échantillonnage est celui qui se présentera lors du Recensement de l'an 2000 pour lequel l'échantillonnage aux fins du suivi est interdit. Examinons d'abord la section A pour un secteur comptant 2 500 habitants. Si l'on ne

$$REQM(T) = \sqrt{(0,02T)^2 + (0,25)(3)TQ}$$

Cette formule est celle sur laquelle se fondent les calculs présentés au tableau 2. Dans le cas d'un chiffre de recensement non corrigé, $REQM(T)$ comprend à la fois les termes de biais et de variance. Dans le cas d'un chiffre de recensement corrigé, le biais relatif est nul, si bien que seul le terme de variance de correction sont eux-mêmes dépourvus de toute forme de variance ou de biais et que l'on peut appliquer uniformément les mêmes facteurs de correction à tous les groupes démographiques.

Par exemple, le tableau 2 montre que pour un total de 500 dans un territoire comptant 2 500 habitants, la $REQM$ est égale à 20 pour un chiffre de recensement non corrigé, et à 17 pour un chiffre corrigé. Dans le cas de la non-correction, la contribution du terme de biais est faible, soit $[(0,02)(500)]^2 = 100$. La contribution du terme de variance de réponse est $(0,25)(3)(500)(0,8) = 300$. Si bien que $REQM = \sqrt{400} = 20$. Dans le cas de la correction du chiffre de recensement, le terme de biais, c'est-à-dire 100, disparaît, si bien que la $REQM = \sqrt{300} = 17$. Cependant, si l'on considère que le terme de biais estimatif comporte à la fois un élément de variance et de biais, la différence entre les résultats corrigés et non corrigés pourrait être faible pour un petit territoire. Comme le total, T , devient de plus en plus grand, le terme de biais prend plus d'importance et la correction élimine une plus grande part de l'erreur.

Le tableau 2 montre que, pour un petit territoire de 2 500 habitants, le gain d'exactitude est nul pour les totaux dont la valeur est faible, mais qu'il se chiffre à 43% pour un territoire de recensement un peu plus grand comptant 10 000 habitants, la réduction de l'erreur est faible jusqu'à un total étudié de 1 000 personnes, pour lequel on observe

Tableau 2
REQM prévue des fréquences estimatives par cellule pour les estimations de population fondées sur des chiffres de recensement corrigés et non corrigés pour le sous-dénombrement

Territoire de 2 500 habitants, REQM fondée sur				Territoire de 10 000 habitants, REQM fondée sur				Territoire de 50 000 habitants, REQM fondée sur			
Fréquence par cellule	Non corrigés	Corrigés	Fréquence par cellule	Non corrigés	Corrigés	Fréquence par cellule	Non corrigés	Corrigés	Fréquence par cellule	Non corrigés	Corrigés
15	3	3	50	6	6	250	15	15	14	22	19
50	6	6	100	9	9	500	22	34	27	65	42
100	9	8	200	13	21	1 000	34	116	58	214	77
500	20	17	500	21	33	2 000	65	250 000	509	97	
750	25	20	1 000	33	26	5 000	116				
1 000	29	21	2 000	53	35	10 000	214				
1 500	37	21	5 000	109	43	25 000	509				

Nota:

Les valeurs sont calculées en supposant que le biais relatif de réponse est égal à 2% si le chiffre de recensement n'est pas corrigé et nul s'il l'est. Dans le cas des chiffres corrigés ainsi que non corrigés de recensement, la variance de réponse est égale au quart de ce que serait la variance d'échantillonnage observée pour un échantillon au 1/4.

un gain d'exactitude de 21%, mais pour un total plus important de 5 000 personnes, le gain est de 61 %. Par conséquent, si nous nous intéressons au nombre d'hommes et de femmes dans un territoire comptant 10 000 habitants, c'est-à-dire un total probablement de l'ordre de la moitié du chiffre de la population, le gain d'exactitude serait important si l'on se servait des chiffres corrigés de recensement. Pour un territoire comptant 50 000 habitants, le terme de biais est le plus important dans l'erreur quadratique moyenne, même pour des totaux de faible valeur comme 1 000 habitants. Ici, le gain d'exactitude est de 21%, mais l'erreur totale n'est jamais inférieure à la variance de réponse. Si l'on corrige les chiffres du recensement, le terme de biais devient nul et le gain d'exactitude est spectaculaire.

L'un des avantages de ce modèle tient au fait qu'il a été mis au point par le Census Bureau bien avant que ne s'échauffe le débat actuel sur la correction des chiffres de recensement. Le Bureau s'en est servi pour détromper les personnes qui pensaient que les valeurs des cellules du recensement n'étaient entachées d'aucune erreur. Il s'en est également servi pour faire admettre aux critiques de la méthode que le fait de ne poser la plupart des questions du recensement qu'à un échantillon de la population n'amoindrirait pas indûment la qualité des données. Un modèle de recensement tel que celui-ci, qui a fait ses preuves, montre la valeur réelle de la correction.

L'exemple qui précède montre que la correction des chiffres de recensement n'augmente pas l'erreur qui entache ces chiffres, même dans le cas des petits territoires de dénombrement et des petites cellules, si l'on suppose que le terme de biais est mesuré sans erreur. Pour les territoires et cellules dont l'effectif est faible, la variance de réponse est le terme dominant de l'erreur quadratique moyenne, mais l'erreur totale n'est jamais inférieure à la variance de réponse. Si l'on corrige les chiffres du recensement, le terme de biais devient nul et le gain d'exactitude est spectaculaire.

L'un des avantages de ce modèle tient au fait qu'il a été mis au point par le Census Bureau bien avant que ne s'échauffe le débat actuel sur la correction des chiffres de recensement. Le Bureau s'en est servi pour détromper les personnes qui pensaient que les valeurs des cellules du recensement n'étaient entachées d'aucune erreur. Il s'en est également servi pour faire admettre aux critiques de la méthode que le fait de ne poser la plupart des questions du recensement qu'à un échantillon de la population n'amoindrirait pas indûment la qualité des données. Un modèle de recensement tel que celui-ci, qui a fait ses

5. AUTRES APPLICATIONS DES OUTILS STATISTIQUES

On peut se servir des outils statistiques pour corriger les données du recensement de façon à tenir compte du sous-dénombrement. Le modèle de REQ_M de Wakseberg-Hanson permet d'estimer la valeur de l'erreur qui entache le chiffre de recensement si l'on suppose que le biais relatif de réponse est de 2% pour l'ensemble du recensement. (L'estimation était de 1,6% pour 1990.) On suppose aussi que la variance de réponse est égale au quart de la variance d'échantillonnage observée pour un échantillon au 1/4, aussi bien pour les données corrigées que non corrigées du recensement. Aujourd'hui, cette estimation pourrait être trop faible, puisque la baisse du taux de retour des questionnaires par la poste a fait augmenter la variance liée aux recenseurs. Cependant, par prudence, nous utiliserons les évaluations de 1960 et de 1970.

Le modèle est celui de l'erreur quadratique simple utilisé fréquemment par le Census Bureau.

$$EQM(T) = Var(T) + B_T^2$$

Supposons que T représente l'effectif d'une cellule ou un territoire où N représente la taille de la population, $T = NP$, où P représente la proportion de la population présentant une caractéristique particulière. B représente le biais qui entache le chiffre de recensement. Donc, on pourrait, par exemple, vouloir déterminer le nombre d'enfants de moins de 10 ans dans un territoire de recensement comptant 2 500 habitants. $N = 2\,500$ et $T = NP$. Maintenant, la variance d'une proportion estimative, p , s'écrit:

$$V(p) = \frac{N-1}{N} \cdot \frac{1}{n} \cdot p \cdot \bar{p}$$

Si nous avons prélevé un échantillon au 1/4, cette expression se réduit à:

$$V(p) = \frac{4}{3} \cdot \frac{1}{n} \cdot p \cdot \bar{p} = \frac{N}{3p\bar{p}}$$

$$V(T) = V(Np) = N^2 V(p) = 3Np\bar{p} = 3T\bar{Q}$$

Le biais relatif $= (0,02)$, si bien que le biais $= 0,02T$. Si nous considérons maintenant un recensement, la variance d'échantillonnage est nulle, mais la variance de réponse est égale au quart de ce que serait la variance d'échantillonnage. Par conséquent,

$$EQM(T) = (0,02T)^2 + (0,25)(3)T\bar{Q}$$

D'autres études ont porté sur de nouveaux moyens d'évaluer le sous-dénombrement, sur la contre-vérification des dossiers pour déterminer l'exactitude des données de recensement et sur l'usage du bureau de poste non seulement pour livrer les questionnaires de recensement, mais aussi pour faire part des adresses manquantes et des formulaires en double aux autorités chargées du recensement. On recourt maintenant fréquemment à l'échantillonnage pour contrôler la qualité des travaux de bureau à grande échelle que nécessite le recensement. Lors des recensements antérieurs, on procédait habituellement à une vérification dépendante en vertu de laquelle le vérificateur examinait le travail du codeur et déterminait si celui-ci attribuait les codes correctement. Le Bureau a introduit volontairement des erreurs et a constaté que le taux d'erreurs non repérées par la vérification dépendante pouvait atteindre 50 %. Cette étude, ainsi que d'autres travaux, ont poussé le Bureau à mettre au point une méthode de vérification indépendante, dans le cadre de laquelle des enregistrements sont attribués à trois codeurs qui ne peuvent voir le travail effectué par les deux autres. Le Bureau applique une «règle de majorité» pour déterminer le meilleur code et se sert des statistiques relatives à ce genre d'erreur pour améliorer le processus et repérer les rendements inférieurs aux normes. L'imputation est un autre outil dont la mise au point a été nécessaire dans le cadre du recensement. Pour respecter le calendrier et le budget, le Bureau a mis au point un système d'imputation «hot-deck» qui suppose que les personnes qui vivent dans le même quartier présentent vraisemblablement de nombreuses caractéristiques communes, comme le niveau de scolarité et le revenu. Une autre méthode d'imputation a également été appliquée en 1970, en 1980 et en 1990 pour traiter un petit ensemble résiduel d'adresses figurant sur la liste d'envoi pour lesquelles on ne savait pas si les logements étaient occupés ou non. Personne ne répondait à la porte et les voisins ne savaient pas si les logements étaient habités. Donc, en se fondant sur un modèle supposant qu'il existe une forte corrélation entre les caractéristiques de ménages voisins, le Bureau a imputé un état d'occupation ou de non-occupation et, pour les logements considérés comme occupés, un nombre d'habitants. En 1980, 762 000 personnes seulement ont été imputées de cette façon, soit 0,003 du chiffre total de recensement, mais elles n'étaient pas réparties uniformément entre les divers États. À cause de cette imputation, l'Indiana a perdu un siège au Congrès au profit de la Floride. Cependant, il convient de souligner que ne rien faire au sujet des unités non classées reviendrait à les désigner par imputation comme étant toute innocentes. Or, on disposait de renseignements indiquant que plus de la moitié de ces unités étaient, en principe, occupées, si bien que les données fondées sur l'imputation étaient plus exactes que celles fondées sur les dénombrements sans imputation.

totalisations et pour la vérification a donné de bons résultats et les évaluations ont montré que, même en ajoutant l'erreur d'échantillonnage, l'erreur globale était plus faible que celle liée aux méthodes de recensement plus anciennes, sans échantillonnage. Ces résultats ont confirmé les enseignements passés de Mahalanobis.

Durant le Recensement de 1950, le Bureau a étudié de façon approfondie l'effet du biais de réponse et de la variance de réponse sur les données de recensement. Waksberg et Hanson ont prétendu qu'il était erroné de croire que, sans échantillonnage, les données de recensement étaient dépourvues d'erreur. En 1950, on a procédé à une expérience pour estimer l'effet des recenseurs sur la qualité des données de recensement. Par la méthode des sous-échantillons superposés proposée par Mahalanobis, on a jumelé des territoires de recensement adjacents et affecté les recenseurs au hasard. Puisque les tâches des recenseurs couvraient le même territoire de recensement, les discordances entre les réponses obtenues par les deux recenseurs ne tenaient pas à des différences entre les territoires de recensement. Le résultat le plus important de l'étude est celui montrant que la variance de réponse d'un recensement complet réalisé par des recenseurs faisant du porte à porte pour recueillir les renseignements est la même que celle observée pour un échantillon au 1/4 (Hanson et Marks 1958). D'après ce résultat et ceux d'études du biais qui entache les réponses à diverses questions du recensement, Waksberg et Hanson ont créé un modèle de recensement où le biais relatif de réponse est de 6% et où la variance de réponse est égale à la variance d'échantillonnage d'un échantillon de ménages au 1/4. Puis, ils se sont servis de ce modèle pour produire le tableau 1 qui montre la grandeur de l'erreur totale qui entache les données de recensement avec et sans échantillonnage.

Tableau 1
Racine de l'erreur quadratique moyenne (REQM) prévue des fréquences estimatives par cellule pour les réponses aux questions individuelles pour un dénombrement complet et pour un échantillon au 1/4 des ménages

Territoire de 2 500 habitants,			Territoire de 10 000 habitants,			Territoire de 50 000 habitants,		
REQM fondée sur			REQM fondée sur			REQM fondée sur		
Fréquence	Dénombrement	Echantillon	Fréquence	Dénombrement	Echantillon	Fréquence	Dénombrement	Echantillon
complet	complet	au 1/4	complet	complet	au 1/4	complet	complet	au 1/4
12	7	10	50	1	20	250	34	46
50	14	19	200	30	40	1 000	85	105
125	22	31	500	52	67	2 500	180	200
500	49	62	2 000	140	160	10 000	620	650
1 250	89	102	5 000	320	330	25 000	1 520	1 530

Note 1 : Les calculs sont faits en supposant un biais relatif de réponse de 6% et une variance de réponse égale à la variance d'échantillonnage d'un échantillon au 1/4.

Note 2 : On recourt, pour évaluer l'exactitude des résultats (fréquences par cellule), à une forme particulière de moyenne des erreurs réelles qui se produiraient, à savoir la racine de l'erreur quadratique moyenne. Une règle de travail utile consisterait à supposer que, pour environ les deux tiers des résultats d'un recensement ou d'une enquête par sondage, l'écart entre le résultat et la fréquence réelle par cellule n'excède pas la REQM.

l'important des questions ont évolué au fil des ans. La section suivante montre comment l'utilisation d'outils statistiques a changé le visage de recensement au cours du siècle qui vient de s'écouler.

4. MISE AU POINT D'OUTILS STATISTIQUES

POUR UN RECENSEMENT

Deux phénomènes sont à l'origine de l'évolution considérable des méthodes utilisées pour exécuter le recensement décennal des États-Unis depuis 1940, à savoir l'utilisation d'ordinateurs et l'application de méthodes statistiques. Dans certains cas, les deux phénomènes ont été complémentaires, comme dans celui du traitement rapide des données en vue de l'imputation de valeurs manquantes par la méthode «hot deck». Bien que les ordinateurs aient révolutionné profondément le recensement, la suite de la discussion portera essentiellement sur les méthodes statistiques.

Selon Waksberg et Hanson (1965), en 1940, le recours à l'échantillonnage avait trois objectifs. Le premier consistait à recueillir des données jugées complémentaires aux questions sur la langue maternelle, la situation d'ancien combattant et la fécondité à un échantillon au 1/20 de la population. Le deuxième concernait la réalisation d'études analytiques nécessitant la transcription et la codification manuelles des données. Pour éviter qu'elles durent trop longtemps, la transcription et la codification n'ont été effectuées que sur un échantillon de questionnaires du recensement. Enfin, le troisième objectif était le contrôle des opérations de bureau à grande échelle, comme la vérification et le codage des données, la perforation des cartes, et ainsi de suite. Avant 1940, chaque questionnaire était soumis à une vérification.

Le progrès suivant a été décrit comme suit par Waksberg et Hanson:

«Un grand pas en avant en ce qui concerne l'échantillonnage dans le cadre d'un recensement a été accompli lors du Recensement de la population et du logement de 1950 en réponse à un changement profond d'attitude à l'égard du rôle de l'échantillonnage. Alors qu'en 1940, on ne considérait l'échantillonnage applicable qu'aux questions d'intérêt complémentaire ou secondaire, en 1950, on a examiné l'éventail complet des activités de recensement pour déterminer, de façon logique, dans quelles circonstances un dénombrement complet était nécessaire et dans les quelles un échantillonnage fournirait des renseignements suffisants.» [Traduction]

Le recours plus fréquent à l'échantillonnage pour déterminer les caractéristiques de la population, pour les

Aujourd'hui, toute personne qui s'occupe régulièrement

du recensement sait que les données sont entachées d'erreur.

En premier lieu, s'il n'y a aucune variabilité d'échantillonnage quand les questions sont posées à tous les membres de la population, on observe une variance de réponse considérable liée aux recenseurs, aux répondants et aux personnes préposées au codage des données. En deuxième lieu,

les réponses à nombre de questions du recensement sont entachées d'un biais, même si la personne est dénombrée correctement. En outre, les dénombrements sont biaisés si

Bureau met sur pied un programme d'évaluation pour chaque recensement, déterminer la grandeur de l'erreur et se

sert de ces données pour essayer d'améliorer la qualité des données de recensement suivant.

L'erreur de recensement influence la qualité des données sur les groupes de personnes. L'erreur systématique due au

sous-dénombrement touche beaucoup plus les minorités et les enfants que d'autres populations (Edmonston et Schulze 1993). Par conséquent, les collectivités dont la

population est en grande partie afro-américaine, hispanique ou amérindienne sont sous-représentées dans les répartitions éventuelles du pouvoir et des ressources monétaires,

tandis que les statistiques calculées d'après les données sur les enfants de moins de 10 ans sont entachées d'une erreur

importante.

Au fil des ans, le Census Bureau a publié les résultats de nombreuses études visant à trouver un juste équilibre entre

le coût et l'exactitude. L'une mentionnée plus haut est celle du recours à l'autodénombrement. Dans les régions faiblement peuplées, la variabilité de réponse, causée principalement par les intervieweurs, est très forte. Comme dans le

cas de l'erreur d'échantillonnage, l'effet diminue à mesure que la taille de la région, donc le nombre de recenseurs qui recueillent les données, augmente. Quand le taux de

réponse par la poste est de l'ordre de 80%, la variabilité de valeur observée pour un échantillon au 1/4 (Ballar 1969).

Donc, les idées généralement répandues au sujet du recensement ne sont pas toujours correctes. En outre, tous les recensements ne sont pas identiques. Depuis le premier

recensement réalisé en 1790, le Census Bureau a procédé à de nombreuses modifications. Le nombre de questions, le genre de questions, les catégories de personnes dénombrées et les endroits où elles sont dénombrées, les personnes qui

procèdent au dénombrement, la méthode utilisée pour affecter des personnes à un domaine géographique, celle utilisée pour traiter les données manquantes et la sélection

(de plus en plus fréquente) d'un échantillon pour poser la

3. DANS L'ESPRIT DES GENS, QU'EST-CE QU'UN RECENSEMENT

Pour la plupart des gens, faire un recensement signifie envoyer des recenseurs sur le terrain pour qu'ils comptent chaque personne. Trois idées concernant les recensements semblent prévaloir. La première est que chaque personne est dénombrée, la deuxième, qu'un recenseur voit chaque personne et la troisième, que le recensement n'est entaché d'aucune erreur. Examinons-les l'une après l'autre.

Il arrive souvent que tous les membres de la population nationale et que le groupe qui doit l'être varie d'un pays à l'autre, ainsi qu'au fil du temps dans un pays particulier. Ainsi, les membres du personnel militaire et les membres de leur famille postés à l'étranger pourraient ou non être dénombrés. Les civils étrangers qui séjournent provisoirement dans le pays à titre de travailleurs saisonniers pourraient ou non être dénombrés. Ces exemples montrent qu'il est essentiel de définir le champ d'observation de tout recensement.

Donc, par définition, certains groupes de personnes ne sont pas censés être dénombrés lors du recensement. Ces groupes sont définis par le Census Bureau. Certaines personnes prennent personnellement, ou en famille, la décision de ne pas être dénombrées lors du recensement. Par le passé, certaines familles ne déclaraient pas les enfants souffrant de certaines maladies ou d'un retard de développement. Les personnes qui ont eu maille à partir avec le système judiciaire pourraient, elles aussi, décider de ne pas être recensées. Il pourrait s'agir de personnes qui sont entrées illégalement au pays, qui cherchent à échapper aux autorités policières ou qui craignent, quelle qu'en soit la raison, les conséquences d'un recensement. En 1990, certaines personnes ont dit ne pas vouloir participer au recensement parce qu'elles estimaient que ce dernier représentait une trop grande intrusion dans leur vie privée. Enfin, certaines personnes ne sont pas recensées par accident plutôt que de façon intentionnelle. Il se pourrait que ces personnes vivent dans des immeubles qui sont manqués par les recenseurs, qu'elles soient sans logis et qu'elles n'aient donc pas été délistées ou qu'elles soient absentes de leur domicile durant la période du recensement. En 1998, de nombreux rapports ont fait état de la difficulté à recenser les personnes qui vivent dans les collectivités avec portait et loge d'entrée. Il se pourrait que certaines de ces personnes ne soient pas recensées à cause de l'excès de zèle des gardiens. Dans certains collectifs, certains groupes de personnes pourraient ne pas être dénombrés faute de cartes correctes ou à jour.

Quelle qu'en soit la raison, la population entière n'est pas, et n'a jamais été, dénombrée lors d'un recensement. Le deuxième mythe qu'il convient de détruire est celui voulant qu'un recenseur rencontre chaque personne et sait quelles personnes devraient être ou non couvertes par le recensement. Cette situation ne s'est jamais produite, même

lors des tous premiers recensements réalisés aux États-Unis, quand les U.S. Marshals s'occupaient du recensement et que le pays était beaucoup plus petit. En fait, les premiers recensements portaient sur les ménages plutôt que sur les personnes. Autrement dit, on ne posait aucune question à des personnes particulières, car on s'intéressait plutôt au nombre de personnes que comptait le ménage, au nombre d'hommes et au nombre de femmes, au nombre appartenant à chaque groupe d'âge, et ainsi de suite. C'est en 1880 qu'a vu le jour la méthode du porte à porte pour recenser la population. Cette méthode est la raison pour laquelle certaines personnes ont commencé à croire que tout le monde recevait la visite d'un recenseur. Pourtant, un seul membre du ménage répondait habituellement pour toute la famille. Le recenseur ne voyait pas les personnes malades, ni celles qui étaient au travail, qui étaient absentes, provisoirement ou qui n'étaient, pour une raison ou une autre, pas présentes dans la pièce lors de sa visite.

Même si le recours à des recenseurs constituait une amélioration par rapport au recensement réalisé par les marshals, les études portant sur les sous-échantillons superposés ont montré que l'utilisation de recenseurs entachait encore d'une erreur considérable les statistiques tirées du recensement. En effet, les recenseurs étaient influencés par leurs propres attentes et par les réponses d'autres personnes de leur district de recensement. En outre, certains ne comprenaient pas les instructions et déclaraient incorrectement les renseignements fournis. Une expérience réalisée lors du Recensement de 1950 a montré que les recenseurs augmentaient considérablement la variance des statistiques tirées du recensement (Hanson et Marks 1958). En effet, la variance due aux recenseurs des statistiques tirées d'un recensement était la même que celle des statistiques fondées sur un échantillon au 1/4. Cette observation est la principale raison pour laquelle le Census Bureau a commencé à appliquer lors du Recensement de 1960 une méthode d'autodénombrement dont on a progressivement étendu l'usage lors des recensements ultérieurs. Aujourd'hui, si un ménage reçoit le questionnaire du recensement par la poste, le retourne d'abord rempli et qu'aucune erreur n'exige une prise de décision, aucun recenseur ne rendra visite au logement de ce ménage.

Le troisième mythe est celui voulant que le dénombrement se fasse sans erreur. Toute personne qui travaille aujourd'hui au recensement sait qu'il n'en est rien, mais d'autres y croient. Le Census Bureau encourage cette croyance, parce qu'il publie des données à l'unité près. Par exemple, le Bureau a publié dans la revue *Statistical Abstract* qu'en 1990 les États-Unis comptaient 248 718 301 habitants.

Même des personnes impliquées de près dans les travaux du recensement ne peuvent imaginer ce dernier comme un processus statistique entaché d'une erreur. Comme cette erreur n'est pas quantifiée et publiée systématiquement avec les chiffres du recensement, certaines personnes ne peuvent croire qu'elle existe. Ainsi, une partie des employés de la

complets lors d'un recensement décennal ou d'une enquête, personnes qui vivent illégalement aux États-Unis doivent, en principe, être couvertes par le recensement, nombre d'entre elles craignent que les autorités gouvernementales se servent des renseignements tirés de certaines statistiques regroupées pour organiser des descentes sur certains groupes.

Le recensement décennal est le dernier exemple que je mentionnerai ici pour illustrer les conflits entre la politique et la statistique. Depuis des décennies, on décrit en détail le sous-dénombrement qui entache le recensement et son effet différentiel sur les populations minoritaires: le Cens Bureau a étudié la question pendant des années et dispose maintenant d'outils et de méthodes statistiques pour représenter les membres non dénombrés de la population dans les totaux du recensement. Pourtant, nombre de politiciens s'opposent à ce «rajustement», parce qu'ils craignent son effet sur la délimitation des districts électoraux. Cependant, l'utilisation des données de recensement dépasse de loin le cadre de la répartition des ressources et de la redéfinition des districts. Ainsi, le débat qui a précédé le Recensement de l'an 2000 a été particulièrement vif.

Puisqu'il existe des circonstances où la politique et la statistique sont en conflit, il est bon de faire un retour en arrière et d'examiner les travaux réalisés par Mahalanobis pour le gouvernement de l'Inde. L'application des méthodes de Mahalanobis a permis au U.S. Census Bureau de mieux comprendre les sources des erreurs qui entachent les données de recensement. Je passerai donc d'abord en revue l'apport de Mahalanobis, puis je retournerai à la discussion sur le recensement afin d'examiner les outils statistiques utilisés à l'heure actuelle et ceux que l'on pourrait utiliser et je conclurai en priant instamment le Congrès et le Census Bureau de continuer, comme il en a été la tradition jusqu'à présent, d'améliorer systématiquement le recensement à l'aide d'outils statistiques.

2. LE LEGS DE MAHALANOBIS

Aujourd'hui, je tiens surtout à souligner son apport important dans le domaine des enquêtes par sondage et de la mesure d'erreurs de toute sorte — erreurs d'observation, erreurs de mesure, erreurs d'échantillonnage, erreurs de transcription, erreurs d'impression. Ses premiers travaux, qui visaient à montrer la part de la variance des statistiques imputable aux intervieweurs, portaient principalement sur les statistiques sur les récoltes (Mahalanobis 1950). Il fut l'un des premiers à déclarer, puis à démontrer, que, dans le cas des données d'enquête, l'erreur globale ne correspond pas uniquement à la variance d'échantillonnage, mais aussi à celle liée à l'élément humain. Un moyen d'étudier ce genre d'erreur consistait à utiliser des sous-échantillons superposés. Selon les mots de Mahalanobis,

«Quand deux (ou plusieurs) échantillons sont tirés d'une même population et étudiés conformément au même plan de sondage, les résultats fournis par les divers échantillons sont tous aussi valides les uns que les autres, même s'ils sont produits par des unités opérationnelles différentes; en outre, les divergences entre les divers ensembles d'estimations donnent une idée directe de la marge d'incertitude.»

Mahalanobis et Lahiri (1961) [Traduction]

Durant les années 40, époque où l'échantillonnage n'était pas encore pleinement accepté, Mahalanobis a montré que les statistiques fondées sur des échantillons étaient au moins comparables à celles fondées sur un recensement, et souvent plus exactes. Il était convaincu, comme nombre d'entre nous le sont aujourd'hui, qu'il est plus facile de contrôler des échantillons qu'un recensement. Il a déclaré (Mahalanobis et Lahiri 1961) que la grandeur des écarts observés lors d'un recensement de la production de jute donnait à penser que les estimations fondées sur un recensement pourraient être inexactes pour les petites régions. La composante aléatoire de l'erreur non due à l'échantillonnage pourrait augmenter l'erreur dans une proportion telle que les résultats obtenus pour une grande région ne diffèreraient peut-être pas de ceux obtenus par sondage. Autrement dit, ce qui vaut pour une grande région ne s'applique pas naturellement aux petites régions.

Le U.S. Census Bureau s'est inspiré des méthodes de Mahalanobis pour approfondir l'étude de la variabilité sous-jacente aux chiffres de recensement.

Le passé en guise de prélude

BARBARA A. BAILAR¹

RÉSUMÉ

Mahalanobis a montré comment se servir de la statistique pour éclairer l'élaboration des politiques gouvernementales. Le US Bureau of the Census s'est inspiré de ses travaux innovateurs pour approfondir l'étude des erreurs de mesure qui entachent les données de recensement et d'enquête. Nombre d'idées fausses sont répandues au sujet des recensements, notamment en ce qui concerne les personnes qu'il faut dénombrer et où il faut le faire. Des erreurs sont commises durant le recensement, notamment des erreurs de couverture. Au fil des ans, le US Bureau of the Census a élaboré des méthodes statistiques, y compris des méthodes d'échantillonnage, en vue d'augmenter l'exactitude des données de recensement et de réduire le fardeau de réponse. Un modèle d'erreur (racine de l'erreur quadratique moyenne) a été mis au point pour estimer les effets conjugués de la variance et du biais sur les données de recensement. Ce modèle est utilisé ici pour étudier les effets conjugués de la variance de réponse, de la correction du biais lié au sous-dénombrement et du recours à l'échantillonnage pour le suivi.

MOTS CLÉS : Recensement; Mahalanobis; modèle d'erreur; échantillonnage dans le cadre du recensement.

1. INTRODUCTION

Peut-être en a-t-il toujours été ainsi – en tant que masse d'informations, la statistique ne soutient pas nécessairement les mesures que souhaitent prendre les politiciens. Certains pays ne publient pas les données de recensement, parce qu'à leurs yeux, la connaissance est synonyme de pouvoir. En revanche, dans notre société, les dirigeants usent du pouvoir de la statistique pour justifier la prise de certaines mesures, pour nous faire part du rendement des activités nationales ou pour faire des comparaisons entre groupes. Nous sommes habitués à lire quotidiennement des statistiques et à nous y fier, même si la plupart d'entre nous se soucient peu de savoir comment elles sont produites, par qui et à quel prix.

Au cours des dernières décennies, nous avons été témoins de nombreux conflits entre la statistique et la politique. Les politiciens se servent volontiers des données sur l'emploi et le chômage, particulièrement en période électorale. Si le taux de chômage est faible, le gouvernement sortant le mentionne et s'en attribue le mérite. Si le taux d'activité montre que de nombreux nouveaux emplois ont été créés, le chiffre est cité également. L'un ou l'autre parti politique peut utiliser les données afin de donner du relief à un argument politique particulier. Les efforts de l'administration Nixon en vue de limiter l'accès à ces données se sont soldés par de nouvelles mesures de protection, si bien que les taux d'emploi et de chômage ne sont diffusés que le premier vendredi de chaque mois par le commissaire du Bureau of Labor Statistics durant la réunion du Joint Economic Committee tenue au Capitole. La définition de la pauvreté fait l'objet d'un débat à l'heure actuelle. À l'époque où Molly Orshansky a inventé la mesure de la pauvreté, aucun des grands systèmes de

groupes pourraient refuser de fournir des renseignements touchant les groupes soit moins bien reconnus. Certains occupations. Les problèmes de protection de la vie privée consulter, d'autre part, sont des sources de grandes préoccupations. La confidentialité des dossiers médicaux, d'une exemple, la confidentialité des dossiers médicaux, d'une

Les problèmes de respect de la vie privée abondent. Par gouvernement. dernière analyse, la deuxième étude a été financée par le traitement fédérale ont été formées parce que des milieux puissants s'opposaient à l'étude de cette question. En objections au financement de ces études par l'administration fédérale ont été formées parce que des milieux puissants s'opposaient à l'étude de cette question. En finant par l'État. La deuxième portait sur le comportement sexuel des adolescents. Dans les deux cas, des échantillon national de personnes de 18 à 59 ans, n'a pas été Gagnon, Laumann et Kotlaia 1994), effectuée auprès d'un aux États-Unis. L'une, intitulée *Sex in America* (Michael, mené deux grandes enquêtes sur le comportement sexuel Research Center (NORC) de l'Université de Chicago a remontent en grande partie à Kinsey. Le National Opinion sexuel. Les renseignements que nous possédons à ce sujet On se sert aujourd'hui d'échantillons probabilistes de la transfert, les pauvres seraient mieux lotis qu'auparavant. mesure les avantages médicaux et d'autres paiements de Cette administration a soutenu qu'en incluant dans la jugée davantage au service des riches que des pauvres. l'administration Reagan s'en sont servis pour illustrer le examen de près les chiffres sur la pauvreté. Les critiques de 30 ans. Cependant, chaque administration politique pauvre n'a plus la même signification aujourd'hui qu'il y A cause des revenus gagnés ou des prestations versées, la paiements de transfert mis en place aujourd'hui n'existent.

- SÄRNDA, C.-E., SWENSSON, B., et WRETMAN, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- SCHUCANCY, W.R., MINTON, P.D., et SHANNON, B.S. (1972). A survey of statistical packages. *Computing Surveys*, 4, 65-79.
- SEDRANSKY, J. (1965). A double sampling scheme for analytical surveys. *Journal of the American Statistical Association*, 60, 985-1004.
- SHAH, B.V. (1978). SUDAN: Survey data analysis software. *Proceedings of the Statistical Computing Section, American Statistical Association*.
- SHAH, B.V. (1984). Software for survey data analysis. *American Statistician*, 38, 68-69.
- SIMPSON, H.R. (1961). The analysis of survey data on an electronic computer. *Journal of the Royal Statistical Society, A*, 124, 219-226.
- SKINNER, C.J., HOLT, D., et SMITH, T.M.F. (1989). *Analysis of Complex Surveys*. New York: Wiley.
- STAFFORD, J.E., et ANDREWS, D.F. (1993). A symbolic algorithm for studying adjustments to the profile likelihood. *Biometrika*, 80, 715-730.
- STAFFORD, J.E., et BELLHOUSE, D.R. (1997). Une algèbre informatique pour la théorie des enquêtes par échantillonnage. *Techniques d'enquête*, 23, 3-11.
- WILLCOX, W.F. (1926). The past and future developments of vital statistics in the United States I: John Shaw Billings and federal vital statistics. *Journal of the American Statistical Association*, 21, 257-266.
- WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.
- WOODRUFF, R.S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 66, 411-414.
- WORTON, D.A. (1998). *The Dominion Bureau of Statistics: A History of Canada's Central Statistical Office and Its Antecedents, 1841-1972*. Montréal et Kingston: McGill-Queen's University Press.
- YATES, F. (1960). *Sampling Methods for Censuses and Survey*. (3ième édition). London: Griffin.
- YATES, F. (1973). The analysis of surveys on computers - features of the Rothamsted Survey Program. *Applied Statistics*, 22, 161-171.
- YATES, F., et SIMPSON, H.R. (1960). A general program for the analysis of surveys. *Computer Journal*, 3, 136-140.

- EFFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia: Society for Industrial and Applied Mathematics.
- ESTEVAO, V., HIDIROGLOU, M.A., et SÄRNDAAL, C.-E. (1995). Methodological principles for a generalized estimation system at Statistics Canada. *Journal of Official Statistics*, 11, 181-204.
- FAN, C.T., MÜLLER, M.E., et REZUCHA, I. (1962). Development of sampling plans by using sequential (item by item) selection techniques and digital computers. *Journal of the American Statistical Association*, 57, 387-402.
- FELLEGI, I.P. (1963). Sampling with varying probabilities and without replacement: rotating and non-rotating samples. *Journal of the American Statistical Association*, 58, 183-201.
- FELLEGI, I.P., GRAY, G.B., et PLATEK, R. (1967). The new design of the Canadian Labour Force Survey. *Journal of the American Statistical Association*, 62, 421 - 453.
- FULLER, W.A. (1975). Regression analysis for sample survey. *Sanhva*, C, 37, 117-132.
- GILLIES, D. (1992). *Revolutions in Mathematics*. Oxford: Clarendon Press.
- GODAMBE, V.P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society, B*, 17, 269-278.
- HANSEN, M.H. (1987). Some history and reminiscences on survey sampling. *Statistical Science*, 2, 180-190.
- HANSEN, M.H., et HURWITZ, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, 333-362.
- HANSEN, M.H., HURWITZ, W.N., NISSELSOON, H., et STERNBERG, J. (1955). The redesign of the Current Population Survey. *Journal of the American Statistical Association*, 50, 701-719.
- HARTLEY, H.O. (1946). The application of some commercial calculating machines to certain statistical calculations. Supplement to *Journal of the Royal Statistical Society*, 8, 154-183.
- HIDIROGLOU, M.A., FULLER, W.A., et HICKMAN, R.D. (1980). *SUPER CARP*. Ames: Iowa State U.P.
- HOLLERITH, H. (1894). The electrical tabulating machine. *Journal of the Royal Statistical Society*, 57, 678-689.
- HOOKER, R.H. (1894). Modes of census-taking in the British Dominions. *Journal of the Royal Statistical Society*, 57, 289-368.
- HORVITZ, D.G., et THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- KAIER, A.N. (1895/6). Observations et expériences concernant des dénombrements représentatifs. *Bulletin de l'Institut Internationale de Statistique*, 9, 176-183.
- KAIER, A.N. (1897). *The Representative Method of Statistical Surveys* (1976, traduction anglaise du norvégien). Oslo: Central Bureau of Statistics of Norway.
- KAIER, A.N. (1905). Discours avec discussion. *Bulletin de l'Institut Internationale de Statistique*, 14, 119-134.
- KISH, L. (1957). Confidence intervals for clustered samples. *American Sociological Review*, 22, 154-165.
- KISH, L., et FRANKEL, M.R. (1970). Balance repeated replication for standard errors. *Journal of the American Statistical Association*, 65, 1071-1094.
- KISH, L., et FRANKEL, M.R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society B*, 36, 1-37.
- KLEIN, L.R., et MORGAN, J.N. (1951). Results of alternative statistical treatments of sample survey data. *Journal of the American Statistical Association*, 46, 442-460.
- KONIJN, H.S. (1962). Regression analysis in sample surveys. *Journal of the American Statistical Association*, 57, 590-606.
- KRUSKAL, W., et MOSTELLER, F. (1980). Representative sampling, IV: the history of the concept in statistics 1895 - 1939. *Revue Internationale de Statistique*, 48, 169-195.
- LANSING, J.B., et MORGAN, J.N. (1971). *Economic Survey Methods*. Ann Arbor: Survey Research Center.
- MAHALANOBIS, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, 109, 325-378.
- MANDEVILLE, J.P. (1946). Improvements in methods of census taking and survey analysis. *Journal of the Royal Statistical Society*, 109, 111-129.
- MCCARTHY, P.J. (1969). Pseudo-replication: half samples. *Revue de l'Institut Internationale de Statistique*, 37, 239-264.
- MURTHY, M.N. (1967). *Sampling Theory and Methods*. Calcutta: Statistical Publishing Society.
- NEYMAN, J. (1934). On the two different aspects of the representative method: stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-625.
- PFEFFERMAN, D. (1993). The role of sampling weights when modelling survey data. *Revue Internationale de Statistique*, 61, 317-337.
- PORTER, R.D. (1973). On the use of survey sample weights in the linear model. *Annals of Economic and Social Measurement*, 2, 141-158.
- RAO, J.N.K., et BAYLESS, D.L. (1969). An empirical study of stabilities of estimators and variance estimators in unequal probability sampling of two units per stratum. *Journal of the American Statistical Association*, 64, 540-559.
- RAO, J.N.K., et SCOTT, A.J. (1981). The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76, 221-230.
- RAO, J.N.K., et SCOTT, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *Annals of Statistics*, 12, 46-60.
- RAO, J.N.K., et THOMAS, D.R. (1988). The analysis of cross-classification data from complex sample surveys. *Sociology Methodology*, 18, 213-269.
- RAO, J.N.K., et WU, C.F.J. (1987). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 321-241.

- généralisée de même que plusieurs autres estimateurs ont été programmés dans le SGE, un système généralisé d'estimation conçu par Statistique Canada. Ce logiciel articulé sur SAS s'applique au volet descriptif plutôt qu'au volet analytique des enquêtes. Estevao, Hidroglou et Sâmdal (1995) le décrivent. Il s'agit d'un logiciel qui peut facilement s'adapter dans les bonnes conditions.
- ## 8. CONCLUSION
- Les développements dans le domaine de la recherche sur l'échantillonnage sont inextricablement liés aux méthodes de calcul et aux méthodes informatiques. L'orientation de la recherche sera guidée en partie par les développements en informatique. Ce que l'on peut s'attendre dans le futur immédiat sur le plan informatique ce sont une vitesse et une capacité de mémoire supérieures qui permettent aux logiciels d'augmenter et d'englober un plus grand nombre d'aspects. Les pratiques généralement acceptables en matière d'estimation d'enquête et l'analyse des données d'enquête seront déterminées par le contenu des logiciels généralement disponibles pour l'échantillonnage d'enquête. Les nouvelles méthodes de recherche continueront de s'appuyer de plus en plus sur l'informatique. Un autre développement qui était prévisible est l'explosion d'Internet. Par suite de cette explosion, plusieurs ensembles de données d'enquête complets sont maintenant facilement disponibles sur le Web. Il se pourrait que bientôt la norme consiste à mettre à l'essai les nouvelles techniques sur diverses enquêtes réelles avant de les publier.
- ## REMERCIEMENTS
- La présente étude a été possible grâce à une subvention du Conseil de recherches en sciences naturelles et en génie du Canada.
- ## BIBLIOGRAPHIE
- ANDREWS, D.F., et STAFFORD, J.E. (1993). Tools for symbolic computation of asymptotic expansions. *Journal of the Royal Statistical Society, B*, 55, 613-628.
- BAINES, J.A. (1900). On census-taking and its limitations. *Journal of the Royal Statistical Society*, 63, 41-71.
- BAYLESS, D.L., et RAO, J.N.K. (1970). An empirical study of stabilities of estimators and variance estimators in unequal probability sampling ($n = 3$ or 4). *Journal of the American Statistical Association*, 65, 1645-1667.
- BELLHOUSE, D.R. (1985). Computing methods for variance estimation in complex surveys. *Journal of Official Statistics*, 1, 323-329.
- BELLHOUSE, D.R. (1988). A brief history of random sampling. *Handbook of Statistics*. (Eds. C.R. Rao et K.R. Krishnaiah) 6, 1-14. Amsterdam: North-Holland.
- BENJAMIN, B. (1961). The 1961 census of population. *Incorporated Statistician*, 11, 130-143.
- BERGDAL, M., BLACK, O., BOWATER, R., CHAMBERS, R., DAVIES, P., DRAPER, D., ELVERS, E., FULL, S., HOLMES, D., LUNDQVIST, P., LUNDSTRÖM, S., NORDBERG, L., PERRY, J., PONT, M., PRESTWOOD, M., RICHARDSON, I., SKINNER, C., SMITH, P., UNDERWOOD, C., et WILLIAMS, M. (1999). *Model Quality Report in Business Statistics Volume II: Comparison of Variance Estimation Software and Methods*. London: Office of National Statistics.
- BOWLEY, A.L. (1906). Address to the Economic and Statistics Section of the British Association for the Advancement of Science. York. *Journal of the Royal Statistical Society*, 69, 540-558.
- BOWLEY, A.L. (1926). Measurement of the precision attained in sampling. *Bulletin de l'Institut Internationale de Statistique*, 22 (1), 1-62.
- BOWLEY, A.L. (1936). The application of sampling to economic and sociological problems. *Journal of the American Statistical Association*, 31, 464-480.
- BOX, K., et THOMAS, G. (1944). The Wartime Social Survey. *Journal of the Royal Statistical Society*, 107, 151-189.
- BREWER, K.R.W., et HANIF, M. (1983). *Sampling with Unequal Probabilities*. (Lecture Notes in Statistics, Volume 15). New York: Springer-Verlag.
- CEBUZZI, P.E. (1998). *A History of Modern Computing*. Cambridge, Massachusetts: MIT Press.
- COCHRAN, W.G. (1946). Relative accuracy of systematic and stratified random samples for a certain class of populations. *Annals of Mathematical Statistics*, 17, 164-177.
- COCHRAN, W.G. (1963). *Sampling Techniques*. (2ième édition). New York: Wiley.
- COHEN, S.B. (1997). An evaluation of alternative PC-based software packages developed for the analysis of complex survey data. *American Statistician*, 51, 285-292.
- DAY, N. (1971). *Canadian Computer Census 1971*. Toronto: Canadian Information Processing Society.
- DEMING, W.E. (1953). On the distinction between enumerative and analytic surveys. *Journal of the American Statistical Association*, 48, 244-255.
- DEMING, W.E. (1956). On simplifications of sampling designs through replication with equal probabilities and without stages. *Journal of the American Statistical Association*, 51, 24-53.
- DURBIN, J. (1959). A note on the application of Quenouille's method of bias reduction to the estimation of ratios. *Biometrika*, 46, 477-480.
- DYKE, G. (1995). Obituary: Frank Yates. *Journal of the Royal Statistical Society, A*, 158, 333-338.

probablement incorporée aux logiciels. Pour certains de ses dossiers d'échantillonnage à grande diffusion, Statistique Canada fournit des procédures dans le code SAS pour l'estimation des variances à partir de la méthode bootstrap. Ces procédures sont toutefois propres aux enquêtes en question.

7. LES MODÈLES D'ÉCHANTILLONNAGE

Les modèles ont eu, tout à tour, la faveur des spécialistes de l'échantillonnage, puis sont tombés dans l'oubli. Compte tenu des travaux de pionnier de Neyman (1934), le paradigme de la randomisation et la distribution de la randomisation ont été primordiaux jusque dans les années 60. On n'a toutefois pas cessé d'utiliser des modèles pendant les années intermédiaires. Cochran (1946), par exemple, a utilisé des modèles pour étudier certains plans de sondage. Il a pu conclure que l'échantillonnage systématique constituait un plan intéressant dans le cas de certaines structures de population. Le débat des années 60 sur les modèles découle de la remise en cause des fondements de l'échantillonnage lancée par Godambe (1955). Depuis lors, non seulement l'utilisation de modèles est-elle réapparue dans la théorie de l'échantillonnage, mais elle a progressé considérablement.

Depuis les années 60, l'utilisation des modèles en matière d'échantillonnage a pris plusieurs directions. De même, l'utilisation pratique et générale de modèles pour l'estimation et l'analyse d'enquête n'est possible qu'au moyen d'un ordinateur rapide et du logiciel approprié. Pour continuer dans la même veine, j'adopterais une approche très restreinte par rapport aux modèles en liant leur utilisation à la technologie informatique.

Plusieurs techniques liées aux modèles ont été informatisées, soit par le biais d'exemples numériques dans l'illustration de la technique ou par le biais d'études de simulation visant à examiner le fonctionnement de la technique. À l'heure actuelle, une seule approche quant aux modèles a évolué au point qu'un programme général prêt à l'emploi est disponible. Il s'agit de l'approche axée sur les modèles que C.-E. Särndal a adoptée pendant plusieurs années, et qui a entraîné l'estimation de régression généralisée ou GREG. Les travaux sont résumés dans Särndal, Swensson et Wretman (1992). À l'origine, les travaux faisaient suite aux débats sur les fondements de l'échantillonnage. Dans le cadre d'un modèle, on peut, en quelque sorte, calculer un meilleur estimateur d'un paramètre de population finie. Ceux qui font la promotion de l'inférence de la randomisation ont souligné que quand le modèle ne marche pas, l'estimation associée peut ne pas être très juste. Särndal proposait d'obtenir l'estimation à partir du modèle, puis de l'adapter de sorte qu'elle demeure cohérente et donne de bons résultats dans le cadre de la distribution de la randomisation. On essaie ici de tirer le meilleur des deux mondes. L'estimation de régression

qui connaissait le SAS pouvait facilement se familiariser avec la nouvelle procédure, ou le logiciel, ce qui fait que, d'une certaine façon, il était convivial. De plus, le logiciel s'est maintenu à jour par rapport à la recherche sur les enquêtes. Le programme initial comprenait des programmes pour calculer les erreurs-types pour les estimations d'enquête, y compris les moyennes, les totaux, les proportions et les ratios. Le programme a été élargi de manière à inclure l'analyse de régression à la fin des années 70, période où l'on effectuait de la recherche sur la régression dans les enquêtes complexes. Il s'agit maintenant de programmes s'appliquant à l'analyse de régression, la régression logistique, l'analyse des données nominales et l'analyse de survie. Il s'est aussi tenu à jour par rapport aux développements dans les installations de calcul. À l'origine conçu pour un processeur central, le logiciel est maintenant disponible sur PC. Il conserve ses liens avec le SAS, bien que des procédures d'analyse d'enquête soient en cours d'élaboration pour le SAS.

À l'heure actuelle, il existe plusieurs autres programmes pour l'analyse d'enquête. Les plus populaires d'entre eux, en plus de SUDAAN, sont STATA et WesVarPC. Alors que SUDAAN a été lié au SAS, le développement futur de WesVarPC, qui a l'origine été développé par la société de recherche Westat, a été confié à SPSS. Par ailleurs, les programmes d'enquête dans STATA font partie d'un plus vaste logiciel d'analyse statistique. Comme pour les fusions dans le monde des affaires en général, dans la foule de l'intégration des produits et des services, la tendance de l'aventurer pour ce qui est des logiciels d'analyse des données d'enquête est un logiciel statistique général. Le développement et la maintenance des logiciels statistiques, dans le cadre de la recherche sur les enquêtes ou d'un contexte élargi, sont une entreprise qui prend beaucoup de temps et qui nécessite un important investissement en capital. Seule une organisation bien financée peut se permettre de le faire.

SUDAAN, STATA et WesVarPC, de même que les logiciels GES de Statistique Canada et un autre qui s'appelle CLAN, ont récemment fait l'objet d'un examen et d'une évaluation dans Bergdahl, Black, Bowater, Chambers, Davies, Draper, Elvers, Full, Holmes, Lundqvist, Lundström, Nordberg, Perry, Pont, Prestwood, Richardson, Skinner, Underwood et Williams (1999). SUDAAN et STATA ont aussi été évalués par Cohen (1997). Sur trois des logiciels examinés (STATA, SUDAAN et WesVarPC), SUDAAN semble disposer du plus grand nombre d'options. Par exemple, Bergdahl et coll. (1999) soulignent que SUDAAN exécute l'estimation des variances à l'égard de statistiques complexes au moyen de l'une des techniques suivantes: la linéarisation, la méthode jackknife et la méthode BRR. WesVarPC englobe la méthode jackknife et la méthode BRR, tandis que STATA ne s'appuie que sur la linéarisation de Taylor. À ce jour, aucun des logiciels n'estime les variances à partir de la méthode bootstrap. Sous peu cette technologie sera

erreurs dans les variables ou les erreurs de mesure dans les variables indépendantes.

Konijn (1962) a adopté une autre approche par rapport à l'analyse de régression. À partir d'un plan de sondage en

grappes, il a présupé différents modèles de régression linéaire simple au sein de chaque grappe. Les paramètres qui l'intéressait étaient les moyennes pondérées des paramètres de régression avec les poids découlant de la taille des grappes. Il s'agit d'une approche axée sur le modèle parce que ce sont les paramètres du modèle qui sont intéressants et non un paramètre de la population finie. On a relié aux oubliettes l'approche de Konijn pendant plusieurs années. Cette approche axée sur le modèle a toutefois été abordée dans des publications importantes: Pfeffermann (1993) y renvoie à plusieurs reprises.

Quant aux origines de l'analyse d'enquête dans les

sciences sociales, mentionnons les expériences semblables qui ont été faites dans le domaine de l'analyse des données nominales. Depuis les années 60, les publications sociologiques renferment de nombreux exemples d'analyses de données nominales qui ne tiennent pas compte du plan de sondage. Après que Rao et Scott (1981, 1984) ont développé des analyses des tableaux de contingence et de la qualité de l'ajustement pour les enquêtes complexes, Rao et Thomas (1988) ont essayé de promouvoir la méthode parmi les sociologues dans le cadre d'un exposé de synthèse. Quand on jette un coup d'œil aux index des mots-clés d'accès aux citations, on se rend compte que bien que les travaux aient eu une incidence considérable dans les publications statistiques et médicales, ils ont très peu influé sur les publications sociologiques. Cela peut être en partie attribuable au manque de logiciels informatiques. Le logiciel le plus populaire chez les sociologues, le SPSS, n'est pas doté à l'heure actuelle de programmes d'analyse des données d'enquêtes complexes. Cela pose un problème plus aigu: la régression, l'analyse des données nominales et autres techniques qui ont été proposées pour les enquêtes complexes ne sont pas généralement réalisables sans le logiciel approprié. Fuller lui-même a essayé de répondre à ce besoin en élaborant un programme prêt à l'emploi pour

Hickman (1980).

Frank Yates à la Rothamsted Experimental Station a été le premier statisticien à créer un logiciel pour la recherche sur les enquêtes. Ses travaux ont commencé à la fin des années 50 (Yates et Simpson 1960). À l'origine, on concevait des programmes propres à chaque enquête. Puis, on a conçu un programme général au début des années 60 (Simpson 1961). Bien qu'il s'agissait du premier programme de la sorte dans le secteur et qu'on ait pu y accéder pendant des années, il n'a jamais été très populaire. Au

6. LOGICIEL STATISTIQUE POUR LA RECHERCHE SUR LES ENQUÊTES

- (1) Le logiciel n'était pas facile à utiliser. Dans son article nécrologique sur Yates, Dyke (1995) en a parlé. Il a dit:

Yates croyait que l'analyste devait comprendre la théorie pertinente et, du coup, être en mesure de préciser dans les moindres détails ce qu'il voulait. C'est peut-être pour cette raison que le programme n'était pas très facile à utiliser! Mais sa puissance et sa flexibilité, de même que la clarté nette de ses résultats étaient, et sont encore, exceptionnels. (traduction libre)
- (2) Il coûtait trop cher pour ce qu'il faisait. Le produit ne pouvait pas soutenir la compétition de concurrents qui offraient un produit semblable à meilleur prix. Wolter (1985) donne une liste d'un certain nombre de logiciels disponibles au milieu des années 80. À cette époque, le logiciel était deux fois plus cher que SUDAAN et ne pouvait faire que des mises en tableaux, alors que SUDAAN pouvait en plus analyser la régression et estimer les ratios.
- (3) La commercialisation est un facteur important dans le succès d'un produit. Yates semblait plus intéressé à remanier son produit afin de l'améliorer que d'investir dans sa commercialisation.
- (4) En 1985, le seul soutien technique sur lequel on pouvait compter à l'égard du logiciel était un manuel.

Yates n'était pas le seul à disposer d'un logiciel qui ne marchait pas. J'ai eu la même expérience avec le logiciel d'estimation des variances axé sur trois algorithmes trans-versaux que j'ai développé (Bellhouse 1985). Hormis le facteur «dépendances» (mon logiciel était gratuit), mon logiciel constitue un exemple vivant, pour les trois autres raisons, d'un logiciel qui n'a pas marché.

Au début des années 70, il y avait plus de 40 programmes prêts à l'emploi, écrits principalement en FORTRAN, qui effectuaient des analyses statistiques (Schucany, Minton et Shannon 1972). De ces logiciels initiaux, seulement deux sont demeurés populaires sur le marché: le SAS d'abord lancé en 1970 et le SPSS, lancé à la fin des années 60.

SUDAAN, développé par B.V. Shah du Research Triangle Institute (Shah 1978 et 1984) est un logiciel d'enquête qui a su conserver sa prédominance sur le marché pendant plusieurs années. Il est très bien commercialisé, et son développeur en assure pleinement le soutien. Au départ, on y accédait selon une procédure du SAS; maintenant, il s'agit d'un logiciel autonome. Le lien avec le SAS explique probablement en partie son succès initial. Ceux

5. ANALYSE DES DONNÉES D'ENQUÊTE

Alors qu'on a fait des progrès constants et importants dans la recherche sur les problèmes d'estimation d'enquête ou d'enquêtes complètes au XX^e siècle, en 1970, on s'était très peu penché sur les aspects analytiques des enquêtes. Les expressions «enquêtes énumératives» et «enquêtes analytiques» ont été inventées par Deming en 1950 (Deming 1953). Dans le même article, il donne aussi une définition succincte:

En bref, dans un recensement, on se demande combien? Et dans une enquête analytique, on se demande pourquoi! Il y a une différence entre deux catégories et, dans l'affirmative, dans quelle mesure est-elle importante? (traduction libre)

Cette citation laisse supposer que l'objet des enquêtes analytiques est de comparer les moyennes de domaines. Certainement, tout au long des années 60, ce qu'on entendait par enquête analytique se limitait souvent à cela. Cochran (1963) énonce ce qui suit:

Dans une enquête analytique, on fait des comparaisons entre différents sous-groupes d'une population afin de découvrir s'il y a des différences entre eux qui peuvent nous permettre de formuler ou de vérifier des hypothèses sur les influences à l'œuvre au sein d'une population. (traduction libre)

Dans le cadre de leur discussion des enquêtes analytiques, Yates (1960) ont aussi insisté principalement sur des comparaisons de domaine. Ils ont abordé l'analyse de régression et le problème de l'atténuation, mais n'ont pas traité du problème des poids d'enquête généraux. Skinner, Holt et Smith (1989) attribuent les travaux complètement nouveaux dans le domaine des enquêtes analytiques aux chercheurs en sciences sociales, à Paul Lazarsfeld en particulier. Je m'appuierai sur la théorie de l'analyse de régression dans les enquêtes complexes pour illustrer les liens avec les sciences sociales, dans le cas qui nous occupe, l'économie.

L'une des premières études à avoir tenu compte des poids d'enquête dans l'analyse de régression est celle de Klein et Morgan (1951). À cette époque, les deux spécialistes travaillaient à l'université du Michigan; Morgan travaillait au Survey Research Center. Au début de leur document, ils énonçaient ce qui suit:

Le plan de sondage, les méthodes de collecte des données et le comportement économique sous-jacent contribueront tous à la formulation du modèle. L'étude des données recueillies dans les enquêtes sur la consommation nous a convaincus qu'on ne peut pas estimer les rapports économiques simplement en appliquant les méthodes statistiques conventionnelles parce qu'il y a des difficultés fondamentales que nous classons comme suit: (1) la

pondération des observations, (2) l'hétéroscédasticité, (3) la non-linéarité, (4) le choix de concepts économiques alternatifs, (5) les erreurs d'observation. (traduction libre)

Ils ont abordé les quatre premières difficultés fondamentales et ont omis la cinquième. Dans leur analyse d'environ 2 300 réponses à l'enquête sur les finances des consommateurs, qui était un échantillon à multiples degrés, Klein et Morgan ont utilisé les poids d'enquête par le biais de l'estimation par les moindres carrés généralisés des paramètres de régression, mais ont ignoré l'effet de grappe pour ce qui est de l'estimation de la variance. Ils ont constaté que dans de nombreux cas l'utilisation des poids d'enquête avait peu d'incidence sur les estimations des coefficients de régression, mais qu'il y avait une réduction dans la variance estimée à l'égard de l'erreur de modèle. Mentionnons que Klein est allé travailler ailleurs, mais que Morgan est resté au Survey Research Center de l'université du Michigan. Vingt ans plus tard, ce dernier et un autre chercheur (Lansing et Morgan 1971), ont donné un aperçu de l'avant-garde en ce qui concerne l'analyse des données d'enquêtes économiques. Peu avait changé quant à l'incorporation du plan d'enquête à l'analyse. Il en est de même pour d'autres secteurs de la recherche en sciences sociales; dans de nombreux cas, même les poids d'enquête n'avaient pas été utilisés. Porter (1973) a fait plusieurs références au débat qui se poursuit depuis au moins vingt ans à savoir si l'on doit utiliser les poids d'enquête dans l'analyse de régression.

C'est dans ce milieu que Kish, qui a aussi travaillé au Survey Research Center de l'université du Michigan, a à l'origine proposé le concept de l'effet du plan (Kish 1957), qui consiste en la mesure de l'augmentation ou de la diminution de la variance sur l'échantillonnage aléatoire simple dans une enquête avec un plan autre que l'échantillonnage aléatoire simple. Les effets du plan sont devenus un cœur de nombreux aspects de l'analyse des données d'enquêtes complexes. Pour ce qui est de l'analyse de la régression, Kish et Frankel (1970) ont étudié les effets du plan dans l'estimation des coefficients de régression. Ils ont obtenu leurs estimations de la variance selon la méthode BRR. On ne sait pas trop, d'après leur présentation, quels coefficients ils estimaient. Plus tard, Kish et Frankel (1974) ont expliqué plus clairement les paramètres. Précisément, les paramètres de la population finie sont les mêmes que ceux que l'on aurait obtenus selon l'estimation par les moindres carrés des paramètres de régression de la superpopulation quand l'ensemble de la population finie est disponible. L'estimation de ces paramètres est devenue l'une des approches standard quant à l'analyse de régression des enquêtes complexes. Fuller (1975), à partir des approximations des variances de Taylor, a donné au processus d'inférence intégral un fondement théorique solide en fournissant des théorèmes limites pour les estimations. De plus, il a abordé le problème que Klein et Morgan (1951) avaient ignoré: les

On a suivi deux directions différentes pour ce qui est de l'application de la programmation de FORTRAN à l'échantillonnage d'enquête. L'une était celle des bureaux de statistiques ou des centres de recherche par enquête et l'autre, celle des chercheurs universitaires. Le genre de travaux effectués dans chaque direction est fortement en corrélation avec la nouvelle puissance de l'ordinateur et la domination d'IBM (et du coup de FORTRAN) sur le marché. À la fin des années 60, de nombreux établissements disposaient de nouveaux ordinateurs plus puissants: il s'agissait souvent d'un ordinateur de la série 360 d'IBM annoncé pour la première fois en 1964. En outre, le logiciel (FORTRAN en particulier) est demeuré compatible avec les changements et les versions plus puissantes, surtout en ce qui concerne les machines de la série 360 d'IBM (Cernuzzi 1998). Le Bureau fédéral de la statistique a obtenu son premier IBM 360 en 1969, tandis que, par exemple, les universités de Manitoba, de Toronto et de Waterloo ont obtenu leurs premières machines en 1966-1967 (Day 1971). Dans les bureaux et les centres de recherche, on a informatisé diverses formules et procédures nécessaires à la conception et à l'analyse des enquêtes. Par exemple, Fellagi, Gray et Platek (1967) signalaient que quand on a remanié l'Enquête sur la population active au Canada en 1964-1965, la sélection de l'échantillon selon la méthode d'échantillonnage avec probabilités inégales de Fellegi (1963) a été codée dans un programme FORTRAN. Kish et Frankel (1970), du Survey Research Center de l'université du Michigan, ont signalé qu'ils avaient utilisé le code de FORTRAN pour obtenir des estimations de la variance à l'égard de diverses statistiques, y compris les coefficients de régression au moyen de la méthode BRR. Au milieu des années 60, des chercheurs universitaires ont commencé à se servir de l'ordinateur par le biais du programme FORTRAN pour étudier, de manière numérique ou empirique, la théorie de l'échantillonnage qu'eux ou d'autres avaient conçue. L'un des premiers étaient Sedransk (1965) qui a mené certaines comparaisons d'efficacité en FORTRAN sur un ordinateur IBM 7074 (commercialisé par IBM en 1964) pour un plan d'échantillonnage double. En particulier, on a fait des comparaisons d'efficacité entre les valeurs optimales en ce qui concerne les tailles d'échantillon de la première et de la deuxième phases et une approximation des valeurs optimales. Pour effectuer les calculs, il fallait prendre les valeurs prévues sur une distribution trinomiale à laquelle on avait imposé plusieurs conditions. Dans ce cas-là, on s'est servi de l'ordinateur pour comparer sur le plan numérique les méthodes exactes et les méthodes plan numériques. À la fin de la décennie, un nouveau genre de procédé de recherche axé sur l'ordinateur a émergé. Rao et Bayless (1969), ainsi que Bayless et Rao (1970) ont comparé plusieurs scénarios d'échantillonnage avec probabilités inégales en générant tous les échantillons possibles et en calculant l'erreur quadratique moyenne de la

population finie pour plusieurs populations réelles et fabriquées. Il est par la suite devenu la norme de mener de vastes études empiriques à l'égard de tout estimateur ou plan nouvellement envisagé.

La technologie informatique a changé de manière remarquable ces 30 dernières années. Les ordinateurs modernes sont beaucoup plus rapides, plus petits et disposent d'une capacité de mémoire supérieure. L'augmentation régulière de la puissance de calcul et la disponibilité de langages de programmation standard ont permis aux chercheurs d'approfondir l'analyse des données d'enquête. Ce changement technologique se reflète dans les faits nouveaux concernant la théorie de l'échantillonnage pour ce qui est de l'estimation de la variance. Des années 60 aux années 80, il y a eu trois approches fondamentales en informatique quant à l'estimation de la variance des statistiques d'enquêtes complexes: la linéarisation de Taylor (voir Woodruff 1971, pour les premières références au sujet de son utilisation), la méthode du jackknife (d'abord proposée pour l'échantillonnage par Durbin, 1959) et la méthode BRR (McCarthy 1969). L'augmentation de la puissance de calcul a permis l'apparition d'une nouvelle technique, la méthode bootstrap d'Efron (1982) pour l'estimation de la variance. Cette nouvelle technique statistique, contenue sur la population du développement des postes de travail en réseau RISC (Reduced Instruction Set Computing) exploitées à partir d'un système d'exploitation UNIX, est hautement dévouée de calculs. Au cours des années 80, les postes de travail RISC ont progressivement remplacé la plupart des processeurs centraux dans les organismes de recherche. Vers la fin de cette transition, Rao et Wu (1987) ont élargi le champ d'application de la méthode bootstrap à l'estimation de la variance pour les statistiques lissées dans le cadre des plans d'échantillonnage à plusieurs degrés stratifié.

Les logiciels d'algèbre sont les plus récents progrès à avoir un effet sur la recherche en statistique. Même si ceux-ci existent depuis un bon moment, ce n'est que durant la dernière décennie que, grâce à leurs progrès, ils sont accessibles à de nombreux chercheurs. Grâce aux programmes d'algèbre informatisés, de nombreuses manipulations complexes peuvent être effectuées automatiquement d'erreur. Comme pour plusieurs autres branches de la statistique, de nombreuses manipulations algébriques dans la théorie de l'échantillonnage sont liées aux algorithmes qui génèrent les partitions. Selon les algorithmes informatisés créés par Andrews et Stafford (1993), de même que Stafford et Andrews (1993), Stafford et Bellhouse (1997) ont appliqué les techniques d'algèbre informatisées à la théorie d'échantillonnage d'enquête. À partir de leurs techniques, la plupart des résultats de la théorie d'échantillonnage classique, déjà mentionnés dans des publications ou à obtenir, peuvent être calculés automatiquement.

Sans être à la pointe du développement informatique comme dans le cas des machines Hollerith, le Bureau of the Census des États-Unis a joué un rôle déterminant dans le développement commercial initial du calculateur numérique. Non seulement le Bureau a reçu le premier UNIVAC produit, mais certains de ses employés ont participé aux décisions conceptuelles quant à sa construction (Ceruzzi 1998 et Hansen 1987). L'ordinateur a été livré en mars 1951 et a servi au traitement des données du recensement de 1950. Il a fonctionné jour et nuit jusqu'à ce que la tâche fut terminée. Une fois le travail terminé, l'ordinateur a servi à d'autres recensements et enquêtes, y compris l'enquête sur l'état de la population. La technologie rattrapait maintenant la théorie; l'ordinateur permettait de mieux calculer les estimations de la variance. Il a aussi ouvert de nouvelles possibilités, en particulier l'imputation de valeurs manquantes. Quant aux estimations de la variance, Hansen, Hurwitz, Nisselson et Stenberg (1955) font le commentaire suivant:

Avant d'avoir un ordinateur électronique à grande vitesse, l'UNIVAC, on estimait de manière très approximative les variances afin d'éviter les calculs qui prendraient énormément de temps avec l'équipement disponible. Grâce à UNIVAC, on a pu éviter, en grande partie, le recours aux approximations. Cependant, même avec l'ordinateur électronique, on aurait beaucoup de mal à calculer les variances si l'on devait le faire pour chaque élément directement. On continuera d'utiliser des méthodes approximatives, mais celles-ci seront évaluées par des calculs plus exacts que par le passé. (traduction libre)

D'autres organisations statistiques ont emboîté le pas, mais plus lentement. Au Canada, on s'est peut-être fié en partie à l'expérience américaine. Dans un rapport de 1956 au statisticien du Bureau fédéral de la statistique au Canada sur la question du calcul au Bureau of the Census (repris et cité par Worton 1998), on affirmait ce qui suit:

Les gens qui s'y connaissent... ne sont pas tout à fait convaincus que le système UNIVAC leur a donné les résultats auxquels on devrait s'attendre d'un système informatique. Sans aucun doute, UNIVAC leur a causé beaucoup de problèmes – qui, pour la plupart, ne sont probablement pas attribuables à UNIVAC. Des facteurs comme une piètre programmation, une mauvaise analyse du travail, un personnel d'exploitation inexpérimenté, des problèmes de maintenance et même des frictions entre les trois groupes d'exploitation, c'est-à-dire le personnel spécialisé, le groupe des opérations centrales et l'unité électronique centrale, se traduisent dans le rendement du système UNIVAC. (traduction libre)

4. PROGRAMMATION SCIENTIFIQUE

Le Bureau fédéral de la statistique, maintenant appelé Statistique Canada, a obtenu son premier ordinateur en 1960, un IBM 705. L'ordinateur a servi au traitement des données du recensement de 1961. Comme il a déjà été mentionné, les Britanniques ont utilisé un ordinateur appartenant à l'armée pour traiter les données de leur recensement de 1961. À la fin des années 40, Mahalanobis était inscrit sur une liste de personnes intéressées à obtenir un des premiers UNIVAC (Ceruzzi 1998). Cependant, les rapports annuels de l'Indian Statistical Institute publiés dans *Samkhyā* a révèlent que l'institut n'a pas eu d'ordinateur avant 1956, année où il a reçu un HEC-2M. On pouvait maintenant obtenir des estimations de la variance pour les estimations d'enquête des moyennes, des totaux et des proportions à l'égard des enquêtes à grande échelle. L'utilisation élargie de la technologie dépendait maintenant de deux choses – l'accès à un ordinateur, qui coûtait très cher, et l'accès à un logiciel approprié qui permettait d'exécuter les calculs.

Certains genres de recherche et l'application de leurs résultats ne sont possibles qu'avec l'informatique. Ces possibilités ont augmenté non seulement grâce au développement de la puissance de calcul, mais aussi grâce à la facilité d'accès à la puissance de l'ordinateur par le biais des langages de programmation ou des programmes prêts à l'emploi. Pendant plusieurs années, le langage de programmation scientifique le plus populaire était FORTRAN (Formula TRANslation). C'est IBM qui l'a introduit pour son ordinateur 704. Ce qui a popularisé en partie FORTRAN, c'est le développement du compilateur WATFOR (WATERloo FORtran) à l'université de Waterloo en 1965. Ce compilateur populaire, qui servait à des fins d'enseignement, combiné à la domination d'IBM sur le marché ont rendu FORTRAN accessible à de nombreux étudiants et par la suite aux chercheurs (Ceruzzi 1998). Dans un rapport sur le développement de ses propres programmes informatiques pour la recherche par enquête, Yates (1973) montre à quel point FORTRAN était répandu au cours des années 60. Les programmes de Yates pour l'ordinateur à la Rothamsted Experimental Station avaient à l'origine été conçus à la fin des années 50 avec un code se rapportant précisément à l'ordinateur dont il disposait. Au milieu des années 60, le code a été écrit en Extended Mercury Autocode. À la fin des années 60, le code devait être traduit en FORTRAN au moyen d'un outil de traduction automatique; autrement, on ne pouvait pas s'en servir à tout autre emplacement d'ordinateur. C'est dans Fan, Muller et Rezucha (1962) que j'ai noté la première utilisation de FORTRAN pour l'échantillonnage. Ces trois personnes, qui travaillaient toutes pour IBM, ont conçu des algorithmes et un code d'accompagnement FORTRAN pour sélectionner des échantillons aléatoires simples à l'ordinateur.

on soulève le premier malaise qui accompagne toujours les changements technologiques et les avantages éventuels liés au changement. Quant à l'introduction de ces machines, le rapport énonce ce qui suit :

Contrairement à ce que craignaient certaines catégories de travailleurs, à savoir que la machine Hollerith éliminerait en grande partie les calculs manuels, on a constaté que les études nouvelles et détaillées qui ne pouvaient pas être officiellement menées poussaient maintenant à être sans trop de difficultés. Ainsi, la demande de calculateurs formés pour les étapes ultérieures était à la hausse. En plus des projets de routine exécutés de temps en temps, on a mené des études spéciales comme celles sur la solution mécanique des déterminants, la construction de tableaux, l'ajustement des polynômes orthogonaux, etc. (traduction libre)

Aux États-Unis, Deming (1956) a notamment repris l'idée générale et a proposé des méthodes d'échantillonnage répétée. Le Bureau of the Census des États-Unis s'est servi de cette méthode pour estimer la variance. Au Bureau, l'idée a évolué vers la pseudo-répétition ou éventuellement la répétition compensée (méthode BRR) aux fins de l'estimation de la variance (McCarthy 1969).

3. L'AVÈNEMENT DU CALCULATEUR NUMÉRIQUE

Au départ, on a créé le calculateur numérique à des fins militaires durant la Deuxième Guerre mondiale (Cernuzzi 1998). Pendant quelques années après la guerre, le militaire a continué de jouer un rôle déterminant dans la progression du calcul. En 1950, on avait conçu des usages commerciaux pour l'ordinateur : c'est là que pratique et théorie en matière d'échantillonnage ont commencé à se rejoindre. La première génération d'ordinateur à usage civil comprenait l'UNIVAC suivi par la série 700 des IBM. Ces ordinateurs comprenaient des milliers de tubes à vide comme mémoire interne. Les tubes de la machine IBM faisaient environ trois pouces et contenaient 1 024 bits d'information. L'UNIVAC exécutait à 2,25 Mhz et pouvait effectuer 465 multiplications par seconde. Dans le cas des deux machines, les données étaient saisies au moyen de cartes perforées et les données étaient conservées sur des bandes magnétiques plutôt que sur des cartes perforées. Le recensement de 1961 au Royaume-Uni souligne la continuité du rôle prépondérant du militaire dans le calcul à ce moment-là. Les données du recensement ont été traitées sur un ordinateur IBM 705 (Benjamin 1961). L'ordinateur appartenait au War Office et a été utilisé par la Royal Army Pay Corps. Les travailleurs au recensement pouvaient utiliser l'ordinateur quand l'armée ne s'en servait pas. L'information était d'abord saisie sur des cartes que l'on perforait dans un endroit, puis transférée dans l'ordinateur dans un autre endroit.

marche en tournant une manivelle, comme la Brunsvig utilisée par Pearson et la Millionnaire utilisée par Fisher. Comme la main-d'œuvre requise pour l'analyse augmentait considérablement selon la taille de l'échantillon, on calculait parfois des erreurs-types et, le cas échéant, on appliquait parfois les formules appropriées. Bowley (1936) décrit une situation typique où les calculs d'erreurs-types sont rares :

De façon générale, la mise en tableaux est un emploi plat et ennuyeux, mais l'observation des entrées qui s'accumulent dans un tableau croisé et de la croissance de la continuité émergeant du caractère aléatoire présente un certain intérêt. Quand les résultats prennent la forme d'une courbe de fréquence, et surtout si nous avons des raisons de nous attendre à retrouver une courbe normale et que nous la trouvons, nous avons de bonnes raisons de supposer que nous avons mesuré avec satisfaction une entité réelle. Par conséquent, la distribution des changements de prix ou leurs logarithmes sur une échelle normale sont généralement considérablement validité d'un indice. Dans de tels cas, le calcul de l'erreur-type est raisonnable. (traduction libre)

Box et Thomas (1944) décrivent une enquête d'environ 4 500 répondants stratifiés selon l'industrie au sein de laquelle ils travaillaient. Les erreurs-types, lorsque présentes, ont été calculées au moyen de la formule qui s'applique à l'échantillonnage aléatoire simple. Une décennie plus tard, Deming (1956) constate ce qui suit :

Bien que par définition tout échantillon aléatoire permette d'indiquer une erreur-type valide, c'est un fait que les résultats des échantillons aléatoires ont trop souvent semblé par le passé n'avoir aucune erreur-type en raison de la quantité incroyablement calculs. (traduction libre)

C'est dans ce contexte que Mahalanobis (1946) a suggéré la technique de sous-échantillons superposés. Cette technique, que Mahalanobis a élaboré à l'Indian Statistical Institute durant les années 30 (Murthy 1967 et Deming 1956) est très simple : on sélectionne au moins deux sous-échantillons indépendants conformément au même plan de sondage. L'écart entre les estimations des sous-échantillons du total de la population donne une estimation non biaisée de la variance de l'estimation définitive du total. Sur le plan du calcul, la méthode comporte des avantages distincts, dans le contexte des cartes perforées, qui permettent d'obtenir plus facilement les sommes que les variances. Dans le cas de sous-échantillons superposés, le principal effort de calcul consiste à trouver les estimations des sous-échantillons qui ne sont fondées que sur les sommes. L'Indian Statistical Institute a obtenu sa première machine Hollerith en 1944. Avant cela, on effectuait manuellement les mises en tableaux et les autres calculs. Dans le rapport annuel de l'institut pour 1945-1946 publié dans *Sankhyā*,

2. LES DÉBUTS: LA PREMIÈRE MORTIE DU

XX^e SIÈCLE

Les deux premiers progrès décisifs en ce qui concerne l'échantillonnage d'enquête, l'un lié à la formulation d'un concept statistique et l'autre au développement de la technologie, ont eu lieu à la fin du XIX^e siècle. Au départ, on s'est opposé ou on s'est montré indifférent aux deux progrès, à l'idée bien plus qu'à la technologie. Les deux se sont toutefois imposés: on les a développés plus tard. Il s'agit des progrès suivants: (1) l'adoption d'une «méthode représentative» de l'échantillonnage de Kaier (1895/96, 1897, 1905) en vu d'un dénombrement total dans les relevés démographiques, et (2) le développement des machines à cartes perforées pour le traitement des données par Hollerith (1894). Les deux progrès décisifs étaient directement liés aux travaux d'enquête ou de recensement. C'est la première et dernière fois que les questions d'enquête ou de recensement ont inspiré d'importantes innovations technologiques. Par après, l'échantillonnage d'enquête s'est adapté à la technologie disponible.

Kaier voulait obtenir un échantillon qui serait à peu près comme une population en miniature. L'échantillonnage permettrait d'obtenir des renseignements plus détaillés et de mener un nombre accru d'études spécialisées à une fraction du coût d'un recensement. On s'est d'abord opposé à l'idée: il a fallu plus d'une décennie pour qu'elle soit acceptée.

Herman Hollerith a conçu des machines pour le traitement des données parce que le Bureau of the Census des États-Unis en avait besoin et que le directeur de la statistique de l'état civil du Bureau, John Shaw Billings, a encouragé le projet. Willcox (1926) décrit les événements qui ont conduit au développement:

Pendant qu'on mettait en tableaux les déclarations du dixième recensement [1880] à Washington, Billings se promenait en compagnie d'un collègue dans le bureau où des centaines de commis trans-feraient à la main laborieusement et lentement les éléments d'information des questionnaires aux feuillets de registre. En regardant les commis, il fait remarquer à son collègue qu'il doit bien y avoir une façon mécanique d'exécuter le travail, quelque chose selon le principe du métier jacquard, peut-être, selon lequel les tous sur une carte réguliersent le modèle à tisser. Ses propos ne sont pas tombés dans l'oreille d'un sourd. Son collègue était un jeune ingénieur talentueux qui s'est d'abord consacré à l'idée pouvait être mise en pratique, et que Billings n'avait aucun désir d'en revendiquer la paternité ou de l'exploiter. Il a, par la suite, consacré le reste de sa vie à parfaire l'invention et à garantir son adoption, ce qui lui a rapporté énormément sur le plan personnel et ce qui a bénéficié grandement au monde entier. Je n'ai pas à décrire ni à faire l'éloge des machines de Hollerith. (traduction libre)

Mandeville (1946) donne une description complète de l'établi- sation et de l'utilisation de ces machines. On a appliqué la machine de Hollerith au traitement des données du recensement de 1890 des États-Unis. Alors qu'il a fallu sept ans pour compléter le recensement de 1880, celui de 1890 était terminé au début de 1895. Le Bureau s'est servi de 180 tonnes de cartes qui ont été traitées à une vitesse de 6 900 cartes par journée de six heures et demie. Non seulement la machine permettait d'épargner du temps, mais elle réduisait considérablement les erreurs de tabulation. Le traitement des données du Recensement de 1891 du Canada s'est fait au moyen des machines à cartes perforées. On ne s'en est toutefois pas servi très tôt au Royaume-Uni et dans le reste de l'Empire britannique. On estimait que le niveau de détail requis ne justifiait pas l'utilisation de la machine de Hollerith, car le temps que permettait d'épargner la machine serait compensé par le temps nécessaire à la perforation des cartes (Hooker 1894). Dans un document sur les recensements, Baines (1900) a exprimé une préférence pour la mise en tableaux manuelle, en particulier quand la main-d'œuvre ne coûte pas cher. Malgré les premières hésitations, la machine de Hollerith a continué de s'améliorer et son utilisation dans le domaine de la statistique était largement répandue au milieu du siècle. Hartley (1946) a démontré l'utilisation la plus sophistiquée de ces machines à cartes perforées aux fins de l'analyse statistique. Il s'agissait notamment du calcul des moyennes mobiles et des corrélations sérielles, de même que la solution aux équations simultanées dans les machines de Hollerith.

Après ces innovations presque simultanées et non liées sur le plan des idées et de la technologie, la théorie a supplanté la pratique les cinquante ou soixante années suivantes. Les développements théoriques dans le domaine de l'échantillonnage se sont poursuivis durant la première moitié du siècle. Pour ajouter aux discussions sur l'orientation que devait adopter la «méthode représentative», Bowley (1926) a établi une monographie décrivant tous les résultats théoriques connus en matière d'échantillonnage concernant la sélection au hasard et la sélection raisonnée. De plus, il a élaboré la théorie de l'échantillonnage stratifié selon une répartition proportionnelle. Le triomphe de la randomisation sur la sélection raisonnée est attribuable à Neyman (1934) qui a montré pourquoi la randomisation permettait de mieux régler les problèmes d'échantillonnage que la sélection aussi élaborée des stratégies de répartition optimale pour l'échantillonnage stratifié. Avant le milieu du siècle, le dernier développement majeur, pour ce qui est du plan de sondage avec estimations appropriées et estimations de la variance, était le concept de l'échantillonnage avec probabilités négatives lancé par Hansen et Hurwitz (1943). L'application de ces résultats théoriques a été limitée à billes incandescentes par Hansen et Hurwitz (1943).

Marchant ou Monroe, ou manuelles que l'on mettait en

Évolution de la théorie de l'échantillonnage d'enquête au XX^e siècle et son rapport avec l'informatique

D.R. BELLHOUSE¹

RÉSUMÉ

Le calcul fait partie intégrante de l'analyse statistique en général et de l'échantillonnage d'enquête en particulier. Le genre de l'échantillonnage est retracé en fonction des développements technologiques dans le domaine du calcul. Ce qui est possible en théorie ne peut être mis en pratique qu'avec la bonne technologie de calcul. De même, les nouveaux développements technologiques peuvent susciter de nouveaux domaines d'étude. Il y a cent ans, c'était les besoins des statisticiens qui provoquaient les développements technologiques. Bien que les développements théoriques en matière de théorie de l'échantillonnage aient souvent dépassé les capacités de calcul, la situation actuelle est que les statisticiens d'enquête sont maintenant à la remorque de la technologie de calcul initiée par d'autres plutôt que les catalyseurs à l'avant-plan du changement technologique.

MOTS CLÉS: Analyse des données d'enquête; calculateurs numériques; cartes perforées; programmation scientifique; logiciel statistique; analyse des données d'enquête; estimation d'enquête.

1. INTRODUCTION

On peut aborder l'histoire de l'échantillonnage d'enquête de plusieurs façons. Il y a deux approches très tentantes que nous ne retiendrons toutefois pas aux fins de la présente étude. La première consiste à examiner l'échantillonnage dans le contexte de l'histoire des idées – qui les a formulées, comment et pourquoi on les formule, on en fait la promotion, on les défend, on les abandonne ou on les remplace. Pour ce qui est des personnes, précisons que ce ne sont pas nécessairement celles qui épousent les idées les premières qui sont à l'avant-plan, mais celles qui en font le mieux la promotion ou celles qui les concrétisent le mieux. L'approche de l'histoire des idées a été suivie dans une certaine mesure par Kruskal et Mosteller (1980) ainsi que Bellhouse (1988), qui ont étudié la progression des idées en commençant par l'adhésion à la méthode représentative de Kaier (1897) pour les recensements combinée à l'utilisation de la randomisation dans les enquêtes par Bowley (1906). Toute l'histoire des débats entourant les fondements de l'échantillonnage s'inscrit directement dans cette approche. Depuis que Godambe (1955) a lancé ce débat, on se demande continuellement à quel moment on doit utiliser des modèles quand au plan de sondage et à l'estimation. On peut aussi aborder l'histoire de l'échantillonnage en envisageant la théorie de l'échantillonnage comme une branche des mathématiques, puis en l'intégrant au modèle général d'évolution de la recherche en mathématiques. La chose se complique du fait qu'il existe plusieurs approches quant à l'évolution des mathématiques, comme l'a abordé Gillies (1992). Dans le cadre de l'une de ces approches, on constate que périodiquement certains résultats constituent la base de nouveaux domaines de recherche, tandis que d'autres domaines tombent en désuétude ou semblent

épuisés. Les nouveaux domaines de recherche attirent souvent plusieurs chercheurs talentueux qui tentent de régler les nouveaux problèmes et écartent, du coup, d'autres sujets de recherche possibles. On peut faire des parallèles avec l'échantillonnage. Hansen et Hurwitz (1943) ont obtenu des résultats sur l'échantillonnage avec des probabilités proportionnelles à la taille et avec remise. Puis, Horvitz et Thompson (1952) ont appliqué l'idée à l'échantillonnage sans remise. Le problème fondamental que pose l'échantillonnage avec probabilités inégales sans remise est de trouver un plan de sondage qui donne les probabilités d'inclusion voulues. Plusieurs chercheurs ont étudié la question dont le point culminant est la monographie de Brewer et Hanft (1983). Dernièrement, très peu de documents font la promotion de nouveaux plans de sondage sans remise entraînant des probabilités d'inclusion proportionnelles à la taille d'une variable. Cependant, les statistiques et l'échantillonnage d'enquête ne peuvent pas être mis en équation avec les mathématiques pures. En grande partie, la recherche statistique a son origine dans les problèmes pratiques qui se posent dans l'interprétation et l'analyse des données et non dans les idées abstraites. Compte tenu de l'explosion de la technologie au XX^e siècle, j'ai choisi une autre approche. Il s'agit d'envisager l'histoire de l'échantillonnage au XX^e siècle comme l'histoire de l'effet réciproque entre les idées qui ont été mises en pratique et l'informatique qui a défini les limites de la pratique ou qui a encouragé les idées de nouveaux développements pratiques. On peut catégoriser le développement des techniques d'échantillonnage à l'intersection de deux éléments: l'utilisation des enquêtes à des fins descriptives et analytiques, et le recours ou non à des modèles hypothétiques.

¹ D.R. Bellhouse, Department of Statistical and Actuarial Sciences, University of Western Ontario, London (Ontario) N6A 5B7, Canada.

5. CONCLUSION

Cette dernière section porte sur quelques grandes perspectives de la recherche sur les enquêtes au cours des 10 à 20 prochaines années. La révolution informatique qui a transformé la recherche sur les enquêtes au cours des 25 dernières années se poursuit, et on peut s'attendre à d'autres modifications de plusieurs aspects de la collecte, du traitement et de l'analyse des données d'enquête. De plus, les télécommunications évoluent rapidement, et il est probable que les changements auront un effet sur les modes de collecte des données d'enquête. Il semble probable qu'à l'avenir on aura davantage recours à des plans de sondage en mode mixte, pour bénéficier de nouveaux modes lorsque les répondants y ont accès (par exemple, Internet), les modes traditionnels étant utilisés pour d'autres répondants. L'effet du mode de collecte sur les réponses d'enquête restera donc une préoccupation importante.

En général, il semble probable que la demande de données d'enquête continuera de prendre rapidement de l'ampleur à mesure que les analystes de la politique apprennent à bénéficier des données d'enquête. De plus en plus, on aura besoin d'estimations d'enquête pour de petits domaines, surtout des régions géographiques, les décideurs concevant leurs programmes en fonction de sous-groupes particuliers de la population. À l'heure actuelle, la demande de données d'enquête relève principalement des administrations centrales; à l'avenir, la demande de la part des administrations provinciales et locales pourrait s'accroître. Or le coût des enquêtes est presque le même pour les faibles populations comme pour les grandes. Les administrations locales ne pourront peut-être pas toujours se permettre le coût d'une enquête à moins que l'on ne trouve des méthodes peu coûteuses.

La principale inquiétude liée à l'avenir de la recherche sur les enquêtes est que la volonté des répondants de participer aux enquêtes continue de baisser et que le perfectionnement de la collecte des données ne réussisse pas à neutraliser cet effet. Il en résultera une baisse des taux de

réponse. Cette observation est particulièrement importante pour les enquêtes téléphoniques, dont le taux de non-réponse est déjà élevé. Une hausse appréciable du taux de signaler la fin de la collecte des données par téléphone pour les enquêtes-ménages.

Enfin, la prochaine décennie sera peut-être marquée par l'établissement d'une nouvelle société distincte pour les professionnels de la recherche sur les intérêts de tous les membres de la profession. Puisque l'échantillonnage a dominé l'évolution de la recherche sur les enquêtes au cours des premières années, celle-ci comporte des liens étroits avec les sociétés statistiques. Toutefois, ces liens se fondent surtout sur la statistique des enquêtes. Il existe aussi des liens avec les sociétés qui s'occupent de sondages, d'études de marché et de diverses spécialisations comme la sociologie et la psychologie, surtout pour ce qui est des aspects de la recherche sur les enquêtes autres que l'échantillonnage. On constate également des liens avec les sociétés d'informatique pour les personnes qui s'occupent de calculs d'enquête. Toutefois, il n'existe toujours pas de société qui cherche à réunir toutes les disciplines de la recherche sur les enquêtes. D'ici quelques années, on verra peut-être la création d'une telle société visant à favoriser les échanges entre diverses disciplines en vue de l'avancement du secteur. Un tel tournant n'éliminerait pas le besoin de maintenir les liens actuels entre les spécialistes de la recherche sur les enquêtes et les sociétés statistiques et autres. Ces spécialistes doivent se tenir au courant des principales activités de la recherche sur les enquêtes tout en suivant l'évolution de leur propres disciplines.

REMERCIEMENTS

Je tiens à remercier Joe Waksberg et Dan Levine des précieuses remarques qu'ils ont formulées en vue de la préparation du présent exposé.

démographiques et sanitaires a été lancé peu de temps après la fin de l'EMF, et des enquêtes ont été menées dans une cinquantaine de pays.

L'éducation a fait l'objet de nombreuses enquêtes internationales, y compris la Troisième étude internationale de mathématiques et des sciences (41 pays en 1995) et sa répétition (40 pays en 1999), le Programme international pour le suivi des acquis des élèves (quelque 30 pays en 2000); la deuxième Civics in Education Study (quelque 20 pays en 1999); la Reading Literacy Study de l'IEA (quelque 30 pays en 1991). L'Enquête internationale sur l'alphabétisation des adultes, qui se poursuit, permet de recueillir des données comparables sur l'alphabétisation des adultes dans plusieurs pays. Deux autres exemples sont l'enquête en grappe à indicateurs multiples (UNICEF) et la Social Dimensions of Adjustment Integrated Survey (Banque mondiale). Une activité connexe est la coordination, par Eurostat, des enquêtes de l'Union européenne. Un exemple de collaboration de plusieurs pays à des enquêtes est le programme international d'enquêtes sociales, programme annuel en sciences sociales qui regroupe actuellement 33 pays membres.

Les programmes d'enquêtes internationales ont pris de l'ampleur pour deux raisons distinctes. La première est l'intérêt accru manifesté pour la comparaison des résultats d'enquêtes de plusieurs pays. La deuxième est de venir en aide à des pays en développement et en transition, surtout, leur expertise des enquêtes étant limitée, afin qu'ils puissent mener des enquêtes produisant d'importantes données de planification. On peut s'attendre, pour ces mêmes raisons, à un élargissement appréciable des activités d'enquêtes internationales à l'avenir.

Liens avec les données administratives. La puissance accrue des ordinateurs et la capacité qui en résulte de mener des analyses plus perfectionnées ont suscité une demande de données plus nombreuses sur les unités échantillonnées. Les analystes veulent trouver une réponse à des questions plus complexes que ce n'était le cas par le passé, et certaines données dont ils ont besoin ne se laissent pas facilement recueillir dans le cadre d'une enquête, du moins pour le niveau de qualité exigé. Même si l'on pouvait recueillir les données, la collecte entraînerait un fardeau de réponse excessif. On a donc cherché d'autres sources de données, quitte à lier celles-ci aux réponses d'enquête. Ainsi, les dossiers d'impôt peuvent fournir de précieux profils des gains d'individus échantillonnés au cours d'une période de temps pour laquelle les répondants ne sauraient fournir les données, et les dossiers médicaux peuvent indiquer l'ampleur des dépenses médicales assumées directement par les assureurs alors qu'elle est inconnue des répondants. Ce genre de lien a été largement facilité par le nombre appréciablement accru de documents administratifs sur support électronique.

La possibilité de lier des données de documents administratifs à des données d'enquête sociale suscite un intérêt

considérable depuis quelques années, et certaines enquêtes ont établi ce genre de lien. Toutefois, il faut généralement surmonter des difficultés appréciables pour avoir accès à des données administratives, et la protection de la vie privée des répondants est une question sérieuse. Jusqu'à présent, ces considérations ont beaucoup limité le recours à des liens avec les documents administratifs dans les enquêtes-ménages. Malgré les avantages éventuels appréciables de ce genre de lien, il n'est pas clair dans quelle mesure ces obstacles pourront être contournés.

Par contre, les données administratives sont devenues un élément clé des enquêtes économiques et des recensements; dans un certain nombre de cas, elles ont remplacé les données recueillies antérieurement auprès des répondants. Il en est résulté une baisse appréciable du fardeau de réponse, une amélioration de la qualité des données, une diffusion plus rapide des résultats et une réduction des coûts. **Analyse secondaire.** La puissance accrue des ordinateurs, la multiplication des enquêtes et l'accès à des données d'enquête plus détaillées ont favorisé une croissance marquée de l'analyse secondaire. Les fichiers à grande diffusion sont plus facilement accessibles, parfois dans le cadre d'archives de données, de sorte que les analystes secondaires peuvent mener leurs propres études, les données d'enquête étant ainsi analysées de façon plus approfondie. Une attention accrue doit dès lors être accordée à la protection de la vie privée des répondants; il faut s'assurer que les fichiers de données accessibles aux analystes secondaires ne portent pas atteinte à la confidentialité. Puisque l'analyse secondaire va sans doute continuer de prendre de l'ampleur, il va falloir continuer de trouver des façons de diffuser les données d'enquête tout en protégeant les répondants, sans pour autant limiter sérieusement le genre d'analyse qu'il est possible de mener.

Qualité des enquêtes. Une attention accrue est accordée aux différents aspects de la qualité des enquêtes. Depuis quelques années, divers organismes d'enquête s'intéressent à la qualité du processus d'enquête, appliquant à celui-ci les concepts de la gestion de la qualité totale. On accorde une attention plus grande que par le passé à la qualité prise dans son sens large, qui englobe l'exactitude des estimations, la pertinence, la rapidité de diffusion, l'accessibilité et la rentabilité, de même que dans son sens étroit d'exactitude. Les utilisateurs des estimations d'enquête et les responsables de l'analyse secondaire ont besoin de connaître la qualité globale des données d'enquête, y compris les erreurs d'échantillonnage, la non-réponse et la non-couverture, les erreurs de réponse et les erreurs de traitement. Même si ce besoin est reconnu depuis longtemps, les descriptions courantes de la qualité des enquêtes comportent souvent de sérieuses lacunes. Il semble qu'une attention accrue soit maintenant accordée à cette question. L'introduction de profils de qualité comprenant des rapports pleinement intégrés sur la qualité des données des enquêtes en cours est un phénomène important.

sont reconnus depuis longtemps, et des enquêtes par panel étaient menées au cours des années 1940 et 1950. À cette époque, toutefois, la création d'ensembles de données longitudinales, intégrant des données de divers cycles, était fort complexe. Le fait que l'analyse des enquêtes par panel se faisait souvent largement de façon transversale était une source majeure de critiques de la méthode. De nos jours, les progrès de l'informatique et des techniques d'analyse longitudinale ont modifié la situation considérablement. Néanmoins, la complexité des données longitudinales et, surtout, le problème des données manquantes n'ont pas disparu. À l'heure actuelle, on a largement recours à des méthodes longitudinales d'analyse, même si l'analyse de nombreuses enquêtes par panel se fait toujours largement de façon transversale, trop peu d'attention étant accordée à l'éventail de facettes que leurs données longitudinales pourraient élucider.

La croissance des enquêtes par panel au cours des 20 dernières années a été énorme, englobant tout un choix de sujets, y compris l'éducation, les transitions des travailleurs, la santé et le comportement aux élections. Les enquêtes par panel sur l'économie des ménages, qui s'inspirent de la Panel Study of Income Dynamics lancée en 1968 par l'université du Michigan, ont gagné en popularité et se poursuivent dans plusieurs pays. Mentionnons également l'Enquête sur la dynamique du travail et du revenu de Statistique Canada et la Survey of Income and Program Participation du Bureau of the Census des États-Unis, qui font appel à des méthodes semblables.

Il est tout probable que le recours à des plans de sondage par panel sera encore plus fréquent à l'avenir. Le défi consiste à tirer pleinement profit des données longitudinales résultantes, car le potentiel analytique d'une enquête par panel augmente de façon exponentielle avec le nombre de cycles de collecte de données. De plus, les techniques perfectionnées d'analyse longitudinale élaborées par les biostatisticiens et d'autres spécialistes permettent de mener des analyses beaucoup plus nuancées que par le passé. Il faut de nombreux analystes chevronnés pour assurer l'exploitation intégrale des données recueillies dans une enquête par panel. La croissance de l'analyse secondaire (voir ci-dessous) est prometteuse comme moyen de mieux utiliser les données d'enquête par panel à l'avenir.

Enquêtes internationales. Diverses enquêtes internationales ont vu le jour depuis 25 ans, dont des enquêtes appuyées par des organismes internationaux et des enquêtes nationales indépendantes coordonnées de façon à permettre des comparaisons entre pays. Une importante percée à cet égard a été l'Enquête mondiale sur la fécondité (EMF), comportant des enquêtes menées dans 42 pays en développement et 20 pays développés au cours de la période 1974-1982. On a pu non seulement recueillir des données précieuses sur la fécondité, mais aussi, dans de nombreux pays, offrir une assistance technique pour la recherche sur les enquêtes favorisant la mise sur pied d'une infrastructure pour les sondages. Le programme d'enquêtes

Les notions d'erreur d'enquête totale et de plan de poursuite la recherche.

dont certaines sont imputées. Une stratégie consiste à appliquer des procédures d'imputation multiples à des plans d'échantillonnage complexes, démarche qui fait largement appel à de puissants ordinateurs. On élabore d'autres méthodes dans le cadre du mode d'inférence fondé sur le plan de sondage standard (forcément en fonction d'hypothèses de modèle). Ces méthodes seront peut-être incorporées un jour dans les programmes d'estimation de la variance de l'échantillonnage en vue d'une application aisée.

Erreur d'enquête totale. La discussion a porté jusqu'à présent sur les diverses composantes du processus d'enquête. Toutefois, une enquête bien conçue est la réunion de composantes en un module efficace tenant compte du facteur coût. On a mieux reconnu cet aspect au cours des 25 dernières années, une attention accrue étant accordée au concept de l'erreur d'enquête totale et à celui du plan d'enquête total. Compte tenu de la réduction des ressources, un plan d'enquête représente un compromis entre, par exemple, la taille de l'échantillon, l'ampleur de la conversion de la non-réponse, la longueur du questionnaire et la qualité des données obtenues à l'aide de différents modes de collecte des données. Au moment d'analyser les données d'enquête, il convient d'évaluer la qualité des estimations en fonction de l'erreur d'enquête totale pour toutes les sources, et non pas simplement l'erreur d'échantillonnage. Pour le plan aussi bien que pour l'analyse, il faut des renseignements détaillés sur les différentes sources d'erreur et leur effet sur les estimations d'enquête. De plus, puisque les enquêtes sont des études complexes comportant de nombreux objectifs analytiques, les besoins en données sont considérables. L'abondante documentation sur les erreurs d'enquête provenant de différentes sources facilite l'étude de l'erreur d'enquête totale et du plan d'enquête total en fonction des contraintes budgétaires, mais il y a lieu

4. AUTRES RÉALISATIONS

On trouvera ci-dessous un aperçu de divers domaines de la recherche sur les enquêtes, autres que ceux de nature méthodologique dont il a été question à la section 3, qui ont donné lieu à d'importants progrès au cours des 25 dernières années. Cet aperçu n'est pas exhaustif; on y traite seulement des secteurs qui, selon moi, ont subi un changement majeur.

Enquêtes par panel. Les avantages des données longitudinales obtenues dans le cadre d'enquêtes par panel

les méthodes étaient forcément assez simples. À l'heure actuelle, on utilise largement des méthodes plus complexes, de classe de pondération et de calage, englobant de nombreuses variables auxiliaires, souvent après avoir utilisé des analyses exploratoires pour cerner des variables auxiliaires appropriées.

Échantillonnage. Les principales méthodes liées au plan d'échantillonnage (stratification, échantillonnage à plusieurs degrés, échantillonnage à probabilités inégales, par exemple) ont été élaborées au cours des premières années et décrites dans des manuels publiés dans les années 1950. Les réalisations du dernier quart de siècle ont permis de raffiner et d'élargir ces méthodes, un exemple étant l'échantillonnage à composition aléatoire pour les enquêtes téléphoniques. Encore une fois, c'est la capacité de l'ordinateur de traiter d'énormes volumes de données de recensement et autres grandes bases d'échantillonnage qui a permis aux spécialistes des enquêtes de préparer des plans plus efficaces que par le passé.

Un domaine de recherche, ces dernières années, a été l'étude des méthodes d'échantillonnage pour les populations rares, à l'aide d'une enquête spéciale ou d'une enquête générale avec suréchantillonnage. Il s'agit là d'un volet de l'élargissement de la demande en fonction de résultats pour de nombreux domaines, y compris les plus petits comme les minorités raciales et ethniques, les enfants pauvres, les groupes d'âge et de sexe et les subdivisions géographiques (voir ci-dessous la mention des estimations régionales). Les chercheurs se penchent sur des plans d'échantillonnage et des méthodes de collecte efficaces pour l'échantillonnage de domaines de ce genre dans des situations où il n'existe pas de bases de sondage spéciales. Compte tenu de la croissance ininterrompue de la demande de résultats par domaine, il faut continuer de trouver des façons de sonder des populations rares de façon rentable.

Dans les années 1970, le mode d'inférence fondé sur le plan adopté généralement pour les sondages a été fortement remis en question par des personnes voulant qu'il soit remplacé par les méthodes fondées sur un modèle que l'on utilise ailleurs en statistique. Le débat s'est atténué, et la stratégie fondée sur le plan de sondage est encore utilisée (voir ci-dessous). À cet égard, il y a lieu de clarifier la terminologie: de le début, le mode d'inférence fondé sur le plan de sondage faisait appel à des modèles afin d'accroître la précision des estimations (par exemple les estimations de régression), mais les estimations demeuraient cohérentes pour ce mode d'inférence indépendamment de la validité du modèle. On distingue donc les procédures assistées par modèle des procédures dépendant d'un modèle. La pertinence des estimations dépendant d'un modèle est fonction de la validité du modèle (ou de la robustesse des estimations vis-à-vis des lacunes du modèle). Les récents progrès de l'informaticienne ont favorisé un recours accru à des modèles, et à des modèles plus complexes, dans le cadre du mode d'inférence fondé sur le plan de sondage et assisté par

Ces remarques n'excluent aucunement les méthodes dépendant d'un modèle de la recherche sur les enquêtes. Au contraire, les méthodes de traitement des données manquant décrites ci-dessus dépendent nécessairement d'un modèle. De plus en plus, des méthodes dépendant d'un modèle servent également à préparer des estimations pour de petits domaines (de petites régions géographiques en général). On a besoin de méthodes de ce genre lorsque la taille des échantillons d'un domaine est trop petite (elle est parfois nulle) pour que l'on puisse préparer des estimations fondées sur le plan de sondage qui soient suffisamment précises. Dans une telle situation, on peut produire des estimations régionales en bénéficiant de données d'enquête d'autres secteurs ou d'autres périodes de temps grâce à un modèle statistique qui relie les données d'enquête à d'autres données, généralement administratives. La croissance rapide de programmes sociaux assurant la répartition de fonds parmi de petites entités géographiques a entraîné une demande considérable d'estimations régionales à jour. Par conséquent, l'estimation régionale est devenue un important secteur de recherche depuis quelques années, et le domaine continuera probablement.

L'estimation de la variance pour des plans d'échantillonnage complexes a été un autre secteur important depuis un quart de siècle. Des méthodes fondées sur les approximations de série de Taylor et les méthodes de répétition ont été utilisées dans les années 1960, mais leur application n'était pas chose courante et se limitait surtout aux activités de recherche. Cette situation a changé de façon marquée sous l'effet de la puissance accrue des ordinateurs et de la mise au point de logiciels de calcul des erreurs d'échantillonnage pour des estimations de plan d'échantillonnage complexe (typiquement stratifié à plusieurs degrés). De nos jours, le calcul des erreurs d'échantillonnage est une pratique assez courante dans l'analyse des données d'enquête.

Un phénomène récent important a été l'application de modèles analytiques aux données d'enquête. Le débat se poursuit quant au choix entre un mode d'inférence fondé sur le plan de sondage et un mode d'inférence dépendant d'un modèle. Pour ce qui est de la stratégie fondée sur le plan de sondage, on a pu observer des progrès théoriques à la fois dans l'application de modèles de régression, de modèles catégoriques, de modèles de survie, de modèles à plusieurs niveaux, etc. avec des données d'enquête et dans l'utilisation de logiciels pour le calcul de la variance de ces modèles. Présentement, les analystes d'enquête mènent souvent leurs études exploratoires à l'aide de modules statistiques standard plus souples et ils calculent la variance en fonction du plan de sondage en se servant, durant les étapes finales seulement de leurs analyses, d'un logiciel d'estimation de la variance de l'échantillonnage. À l'avenir, les procédures d'estimation de la variance de l'échantillonnage devraient être davantage intégrées à des logiciels standard.

Un secteur très actif actuellement est le calcul de la variance d'estimations d'enquête fondées sur des réponses

inscrit la réponse à l'aide du clavier de son téléphone. Dans une variante de cette méthode, les réponses orales sont interprétées à l'aide de techniques de reconnaissance de la parole. Le recours à cette méthode pourrait s'accroître à mesure que l'on perfectionne les techniques de reconnais-

sance de la parole.

Un autre progrès récent a été la collecte de données d'enquête par Internet. Une telle façon de procéder est particulièrement intéressante pour certains types d'enquêtes auprès des établissements et pour des enquêtes auprès de personnes ayant accès à Internet et sachant s'en servir. On peut par exemple transmettre le questionnaire par courriel, si l'on connaît l'adresse électronique (comme c'est le cas des employés d'une entreprise ayant son propre réseau). On peut également monter le questionnaire sur un site Web, quitte à ce que le répondant utilise un mot de passe pour y avoir accès. Pour le moment, Internet ne se prête pas aux enquêtes menées auprès du grand public, à cause de la forte proportion de personnes n'y ayant pas facilement accès, de l'absence de plan d'échantillonnage et de taux de réponse probablement faibles. On ne devrait pas succomber à la tentation de recueillir un grand échantillon de réponses Internet à un questionnaire d'enquête en l'absence d'un contrôle approprié. Une telle démarche ne ferait que répéter les erreurs du fameux Literary Digest Poll de 1936.

Données manquantes. Les enquêtes comportent des données manquantes à cause de la non-réponse totale, de la non-réponse partielle et de la non-couverture. Au cours des 25 dernières années, et même auparavant, l'accroissement du taux de non-réponse totale a suscité de plus en plus de préoccupations. Il est difficile de documenter cette tendance, et l'analyse des données de différentes enquêtes ont même suscité des conclusions divergentes quant à l'existence d'une tendance. Et pourtant les responsables des enquêtes sont généralement d'accord qu'il est devenu plus difficile d'obtenir la coopération des gens. Plusieurs explications ont été proposées à cet égard, par exemple le manque de nouveauté des enquêtes, le nombre accru de gens ayant moins de temps libre, la crainte de violence dans les interviews directes, de même que les effets négatifs du télémarketing sur les enquêtes téléphoniques, mais il n'y a pas d'explication définitive. Peu importe les raisons, il faut désormais déployer plus d'efforts que par le passé pour obtenir un taux de réponses élevé. Ainsi, il faut augmenter le nombre d'appels pour communiquer avec les répondants, consacrer plus d'énergie à convertir les gens qui refusent, et faire davantage appel à l'incitation. Au cours de la dernière décennie, on a mené un nombre appréciable d'études expérimentales, lors d'enquêtes par interview directe et téléphonique, afin de vérifier l'effet, sur le taux de réponse, de diverses incitations monétaires et autres; ces études reprenaient dans le contexte d'une interview les analyses menées antérieurement sur les questionnaires postaux.

La non-couverture est un aspect inquiétant des enquêtes téléphoniques, mais elle a suscité moins d'intérêt pour ce qui est des enquêtes par interview directe, et certainement

moins d'attention que le problème de la non-réponse. Et pourtant la non-couverture liée aux enquêtes par interview directe parmi certains segments de la population (les jeunes hommes noirs aux États-Unis par exemple) est parfois élevée. De plus, les unités non couvertes sont peu connues, si ce n'est que l'on peut s'attendre à des différences appréciables relativement aux unités couvertes. C'est là une source d'erreur d'enquête sur laquelle il y aurait lieu de se pencher à l'avenir. La non-couverture peut être particulièrement grave dans une enquête auprès d'une population rare (les adolescents par exemple) lorsque les unités sont échantillonnées dans le cadre d'un processus de sélection de grande envergure. Puisque les unités menées auprès de populations rares suscitent un intérêt accru, ce type de non-couverture mérite une attention particulière.

Il y a vingt-cinq ans, pour la non-réponse partielle, on laissait simplement tomber les cas de l'analyse en question, en calculant par exemple le pourcentage de répartition pour le sous-ensemble de cas comportant des réponses acceptables. On supposait implicitement que le volet non-réponse partielle manquait complètement au hasard (MCAH). On applique toujours cette façon de procéder à de nombreuses enquêtes, mais de plus en plus on a recours à une certaine forme d'imputation pour attribuer des valeurs aux réponses manquantes de façon à tenir compte des réponses fournies aux autres questions d'enquête. Cette stratégie remplace une hypothèse MCAH parfois intenable par une hypothèse de non-réponse manquant au hasard (MAH), voulant que les variables auxiliaires utilisées dans l'imputation. Il est vrai que l'on utilisait il y a vingt-cinq ans des méthodes d'imputation à l'occasion, mais la plupart des documents importants parus sur ce thème sont postérieurs à 1975. Les méthodes courantes s'appuient fortement sur la puissance actuelle des ordinateurs. L'imputation demeure un domaine de recherche actif comportant deux points de convergence: l'élaboration de méthodes d'imputation conservant la structure de covariance de l'ensemble des données d'enquête, sans oublier que presque toutes les variables d'enquête sont exposées à une non-réponse partielle, et le calcul d'estimations de la variance pour des estimations d'enquête fondées sur des données dont certaines sont imputées (voir ci-dessous). La vérification des données, qui est liée étroitement à l'imputation, a également connu des progrès appréciables ces dernières années, la puissance accrue des ordinateurs ayant permis de mettre au point des procédures plus complexes que par le passé. Comme l'imputation, la vérification fait l'objet de beaucoup de recherches, et l'avenir est prometteur.

La puissance accrue des ordinateurs est également un facteur important du développement et de la fréquence d'utilisation des corrections pondérées de la non-réponse et de la non-couverture. Les corrections par classe de pondération pour la non-réponse et la non-couverture (stratification a posteriori) s'appliquaient lorsque l'on utilisait le matériel d'enregistrement pour l'analyse des enquêtes, mais

que l'on peut atteindre grâce à des interviews directes. Or, le risque de biais appréciable associé aux niveaux élevés de non-réponses des enquêtes téléphoniques (pouvant souvent atteindre 40% ou plus, même en présence d'un suivi dynamique) représente un problème sérieux et souvent sous-estimé. Puisqu'il est probable que le taux de non-réponses des enquêtes téléphoniques va augmenter, on peut se poser des questions au sujet du rôle de la collecte téléphonique à l'avenir.

Un progrès important a été réalisé pour ce qui est de la collecte des données grâce à la récente introduction de méthodes assistées par ordinateur, par exemple les interviews sur place assistées par ordinateur (IPAO) et les interviews téléphoniques assistées par ordinateur (TTAO). Ces méthodes prévoient des instructions de type « passez à la séquence prévue du questionnaire, facilitez les insertions à partir de questions antérieures (exemple: si « Pierre » a été inscrit comme nom du fils lors d'une question antérieure, « Pierre » peut faire partie du libellé d'une question subséquente) et rendent possibles des vérifications à mesure que l'interview se poursuit, des corrections étant apportées au besoin. Le fait d'entrer les données directement dans un fichier informatique assure également un traitement plus rapide. L'élaboration de programmes généralisés pour la collecte de type IPAO et TTAO, y compris l'échantillonnage et la planification des interviews, représente un travail complexe. Plusieurs programmes peuvent être utilisés à cette fin. L'avenir nous réserve sans doute des programmes plus souples et des systèmes auteurs plus simples d'application.

Depuis quelques années, un autre type de collecte des données d'enquêtes assistée par ordinateur a vu le jour. Il s'agit de l'auto-interview assistée par ordinateur (AIAO), dont il existe plusieurs variantes: l'AIAO-vidéo (le répondant lit les questions à l'écran et inscrit les réponses au clavier), l'AIAO-audio (le répondant entend les questions grâce à des écouteurs reliés à un ordinateur portable et inscrit les réponses au clavier), l'AIAO-audio-téléphonique (l'interview AIAO-audio se déroule par téléphone, et le répondant appelle l'ordinateur ou est transféré à l'interview par ordinateur lorsque la communication téléphonique a été établie par l'intervieweur). Toutes ces versions de l'AIAO évitent l'interaction répondant-intervieweur des autres méthodes d'interview, et pourraient donc être d'une utilité particulière pour la collecte de données sur des sujets délicats. Au besoin, on pourrait également prévoir des versions autres qu'anglaise. Les variantes audio n'exigent pas que le répondant sache lire. Ces méthodes ne sont apparues que récemment, et on peut en prévoir l'essor à l'avenir.

Il existe actuellement des enquêtes-entreprises menées à l'aide de méthodes AIAO-audio. Un avantage est que le répondant peut composer un numéro de libre-appel à un moment qui lui convient. Il entend alors des questions d'enquête transmises par numérisation de la parole, et

l'intérêt soutenu que l'on manifeste pour les essais préliminaires est attribuable en grande partie à ce mouvement. Un résultat immédiat du mouvement CASM a été la création des « laboratoires cognitifs » dont on se sert beaucoup maintenant pour les essais préliminaires de questionnaires à l'aide de techniques comme la réflexion à haute voix et les questions exploratoires. Les groupes de discussion, associés depuis longtemps à la conception des questionnaires, surtout pour les études de marché, sont eux aussi utilisés beaucoup plus souvent que par le passé. De plus, le codage du comportement est utilisé largement à l'heure actuelle pour les essais préliminaires.

Un phénomène récent connexe a été l'adoption d'une stratégie plus théorique pour la conception des formulaires d'enquête que les répondants doivent remplir. Ce genre de recherche tient compte de théories qui expliquent comment les particuliers abordent un document et comment ils le consultent. Ce domaine important a suscité peu d'intérêt pendant plusieurs années. Les recherches actuelles sont très prometteuses et devraient rendre les formulaires d'enquête plus conviviaux, ce qui devrait favoriser une amélioration de la qualité des données recueillies et des taux de réponse. **Collecte des données.** *Survey Methods* contient deux chapitres qui traitent des méthodes de collecte des données: le premier aborde les questionnaires postaux et le deuxième les interviews directes (un chapitre décrit aussi les documents et l'observation). On ne mentionne que brièvement les interviews téléphoniques, en partie à cause de la faible pénétration du téléphone au Royaume-Uni à cette époque. Toutefois, même aux États-Unis où la pénétration du téléphone était beaucoup plus marquée, en 1975 de nombreux spécialistes des enquêtes avaient de sérieuses réserves au sujet de la collecte téléphonique des données pour les enquêtes-ménages, du moins pour les enquêtes gouvernementales comportant d'importantes répétitions en matière de politique. La situation a beaucoup évolué. De nos jours, de nombreuses enquêtes du gouvernement des États-Unis se font par téléphone.

Une préoccupation suscitée par les enquêtes téléphoniques est la non-couverture des ménages sans téléphone. Puisque la couverture téléphonique se rapproche actuellement de 95% aux États-Unis, la non-couverture des ménages sans téléphone peut être considérée comme acceptable pour les enquêtes auprès de la population générale. Toutefois, un nombre important d'enquêtes se penchent sur des sous-populations dont le taux de couverture téléphonique est moins élevé, par exemple les personnes pauvres; la non-couverture téléphonique est une préoccupation sérieuse pour de tels ménages. La non-réponse est elle aussi inquiétante. Les taux de non-réponse pour les enquêtes téléphoniques sont appréciablement plus élevés que pour des enquêtes comparables fondées sur des interviews directes, et l'écart semble s'élargir. Lorsqu'il faut faire un choix entre des modes de collecte téléphoniques et directs, les économistes considérables réalisées à l'aide des interviews téléphoniques l'emportent souvent sur le taux de réponse plus élevé

des 25 à 30 dernières années. La capacité de traiter et d'analyser les données d'enquête beaucoup plus aisément que par le passé a favorisé l'utilisation de méthodes statistiques plus avancées. Elle a également suscité des demandes beaucoup plus détaillées de la part des utilisateurs des données d'enquête, stimulant la mise au point de méthodes améliorées pour tous les aspects du processus d'enquête.

Le chapitre sur le traitement des données d'enquête dans *Survey Methods* contient une description des cartes perforées qui étaient utilisées largement il y a trente ans pour l'analyse des données d'enquête, de même que du matériel d'enregistrement (compteuses-trieuses et tableaux) et des ordinateurs. À cette époque, les ordinateurs, étaient à la veille de remplacer le matériel d'enregistrement, mais ils n'étaient pas habituellement accessibles aux spécialistes des enquêtes. Les ordinateurs de l'époque étaient de gros appareils centraux, et les cartes perforées étaient l'outil habituel de saisie des données d'enquête. Le nombre et l'envahissement des programmes d'analyse des enquêtes étaient limités. De nos jours, la situation est tout ce changement pour la recherche sur les enquêtes.

C'est dans le contexte de cette explosion informatique qu'il convient d'évaluer les progrès réalisés dans d'autres secteurs de la méthodologie des enquêtes. Le reste de la présente section décrit brièvement les progrès importants qui, à mon avis, ont été réalisés au cours du dernier quart de siècle pour ce qui est de la conception des questionnaires, de l'échantillonnage et de l'erreur d'enquête totale.

Conception des questionnaires. Le rôle essentiel de la conception des questionnaires dans l'obtention de données d'enquête de qualité élevée a été reconnu dès le début. Il est vital que d'excellents travaux ont été menés sur l'amélioration de la conception des questionnaires au cours des années 1960 et 1970, mais le nombre de chercheurs qui se penchaient sur cette tâche très épineuse était très limité. La situation s'est améliorée appréciablement à cause surtout de ce que l'on a appelé le mouvement CASM (aspects cognitifs des méthodes d'enquête). Le mouvement CASM cherche à recruter des spécialistes des sciences cognitives et sociales pour aborder les problèmes épineux de la formulation des questions d'enquête pouvant susciter des réponses appropriées. L'intérêt suscité par ce mouvement a entraîné un nouvel essor dans ce secteur.

Le mouvement CASM n'a pas mis en évidence des solutions toutes faites aux problèmes des erreurs de réponse dans les enquêtes. Il n'aurait pas été raisonnable de s'attendre à tout régler par l'importation de théories existantes de la psychologie cognitive et d'autres disciplines. Le mouvement a permis de multiplier les efforts dans une perspective théorique. De plus, le mouvement CASM a permis d'établir des essais préliminaires plus rigoureux pour les questionnaires d'enquête. Certaines techniques d'essai préliminaire élaborées au cours des 25 dernières années sont survenues indépendamment du mouvement CASM, mais

doivent avoir recouru à des méthodes beaucoup plus perfectionnées que par le passé pour la saisie et le traitement des données. Cette segmentation inévitable des méthodes d'enquête, compte tenu des progrès accomplis, risque de nuire à l'unité de la profession, d'autant plus que les sous-disciplines donnent naissance à leur tour à des sous-secteurs. Vu l'importance d'une collaboration interdisciplinaire pour ce qui est de la recherche sur les enquêtes, il va peut-être falloir trouver à l'avenir des mécanismes favorisant une telle collaboration (voir la section 5).

Tout comme les pays en développement et en transition, les pays développés n'ont pas suffisamment de méthodologues et de statisticiens d'enquête bien formés. Il importe à la fois d'attirer plus de gens à la profession et de prévoir plus de possibilités de formation. Il existe quelques programmes d'études supérieures dans les universités, et certains professeurs se spécialisent dans ce secteur, mais les effectifs sont insuffisants compte tenu des besoins. La collaboration multidisciplinaire nécessaire à la préparation et à l'exécution d'une enquête suppose que la formation comporte un volet multidisciplinaire, afin que les spécialistes puissent communiquer de façon efficace entre eux. De plus, parmi les instructeurs on devrait trouver des personnes ayant une expérience pratique des enquêtes. Compte tenu de ces exigences, il est difficile de mettre sur pied un programme d'études supérieures en méthodes d'enquête dans la plupart des universités. Une autre stratégie a été adoptée dans le cadre du Joint Program in Survey Methodology (JPSM) de l'université du Maryland, mis sur pied grâce à des fonds publics aux États-Unis pour combler le manque de spécialistes d'enquête dûment formés dans l'administration fédérale. Ce programme se fonde sur une collaboration entre deux universités (l'université du Maryland et l'université du Michigan) et un organisme privé de recherche sur les enquêtes (Westat), avec la participation notable de spécialistes en méthodes d'enquête du secteur public, d'autres organisations et d'autres universités. De même, le Department of Social Statistics de l'université de Southampton et l'Office for National Statistics du Royaume-Uni ont récemment élaboré un programme de deuxième cycle universitaire en statistique officielle, avec la participation, pour l'enseignement des méthodes d'enquête et d'autres aspects de la statistique officielle, de statisticiens du secteur public. Ce département d'université collabore également avec un organisme indépendant de recherche sur les enquêtes (le National Centre for Social Research) dans le cadre du Centre for Applied Social Surveys, dont une des activités est d'offrir des cours abrégés en méthodes d'enquête.

3. ÉVOLUTION DES MÉTHODES D'ENQUÊTE

La révolution informatique qui a commencé à influencer les analyses d'enquête dans les années 1960 a joué un rôle dominant dans l'évolution des méthodes d'enquête au cours

traiter un thème de façon globale et de produire des textes bien équilibrés. Il s'agissait de combler les lacunes documentaires suscitées par le fait que les méthodologistes d'enquêtes sont des praticiens qui n'ont guère le temps de publier. Il en est résulté la préparation de recueils publiés sur des thèmes comme les enquêtes par panel, les enquêtes téléphoniques, les enquêtes auprès des entreprises, les erreurs de mesure liées aux enquêtes, la qualité des enquêtes et la collecte assistée par ordinateur.

De nombreuses autres conférences sur les méthodes d'enquête ont eu lieu. Les unes ont été organisées par des organismes publics comme Statistique Canada, le Bureau of the Census des États-Unis et le Federal Committee on Statistical Methodology des États-Unis, fondé lui aussi en 1975. Les autres ont été organisées par des associations professionnelles comme l'AISE et l'Association for Survey Computing. Les actes de ces conférences et ceux de la Statistical Association ont largement favorisé l'accroissement de la documentation sur les méthodes d'enquête. Deux autres aspects de l'évolution de la profession de spécialiste des enquêtes méritent d'être soulignés. Le premier est son caractère international. Les conférences internationales décrites ci-dessus ont donné lieu à des publications d'autres de plusieurs pays. Malgré l'existence de différences culturelles entre pays, dont il faut tenir compte dans la collecte des données, la recherche sur les méthodes d'enquête comporte de nombreux aspects communs à plusieurs pays. De plus, les enquêtes internationales sont plus fréquentes, d'où le besoin de normaliser les procédures d'un pays à l'autre (voir ci-dessous). De façon générale, la coopération internationale en matière de recherche sur les enquêtes va bon train, mais il existe un secteur qui mérite une bien plus grande attention. Tout comme les pays développés, les pays en développement et en transition ont besoin de données statistiques très d'enquêtes. Toutefois, ils n'ont pas toujours les compétences nécessaires. L'AISE, des organismes internationaux comme le Bureau de statistique des Nations Unies, plusieurs bureaux de la statistique du secteur public et d'autres organismes jouent un rôle important dans la formation de spécialistes des enquêtes dans les pays en développement et en transition, mais le soutien prévu actuellement à cet égard est loin de combler les besoins.

L'autre aspect de l'évolution de la profession de spécialiste des enquêtes est son caractère multidisciplinaire. En trouvant sa place parmi les disciplines professionnelles, la recherche sur les enquêtes a donné lieu à des sous-disciplines. Il y a une trentaine d'années, un méthodologiste des enquêtes pouvait s'attendre à couvrir tous les aspects de sa matière, mais cela n'est plus possible au niveau technique et de la plus élevée. La technicité de l'échantillonnage et de l'analyse des enquêtes est très poussée, et les méthodologistes des enquêtes utilisent de plus en plus des théories et des techniques se rapportant à la sociologie, à la psychologie et à l'anthropologie, tandis que les informaticiens

jugés propices. Des chercheurs aventureux ont réussi à repousser les limites conventionnelles quant aux sujets abordés dans les enquêtes. Cette tendance s'est poursuivie au cours des 25 dernières années, de sorte qu'aujourd'hui très peu de sujets sont exclus d'enquêtes fondées sur des échantillons probabilistes valides. Certains nouveaux sujets d'étude sont délicats, par exemple le comportement sexuel et l'utilisation de drogues illicites, et l'application des méthodes d'enquête a exigé l'élaboration de techniques spéciales de collecte des données. D'autres sujets nouveaux ont supposé l'incorporation de méthodes de collecte supplémentaires. Certains nouveaux sujets d'enquête ont supposé l'incorporation de méthodes de collecte supplémentaires. Certains nouveaux sujets d'enquête ont supposé l'incorporation de méthodes de collecte supplémentaires. Certains nouveaux sujets d'enquête ont supposé l'incorporation de méthodes de collecte supplémentaires.

Avant 1975, il n'existait aucune revue à grand tirage spécialisée en méthodes d'enquête. Les articles évalués traitant de méthodes d'enquête étaient publiés dans différentes revues. Les revues de statistique publiaient, et continuent de publier, des articles sur les méthodes d'enquête se rapportant à leur discipline. Cette situation n'était pas idéale, puisqu'il n'y avait aucun débouché naturel pour de bons articles traitant des méthodes de recherche sur les enquêtes, et parce que la documentation était éparpillée. Le lancement de *Techniques d'enquête* en 1975 et du *Journal of Official Statistics* en 1985 a permis de rectifier la situation; ces deux revues sont maintenant bien établies.

Un autre phénomène important a été l'établissement d'associations professionnelles pour les méthodologistes d'enquête. Ainsi, l'Association internationale des statisticiens d'enquêtes (AISE) a été fondée en 1975 à titre de section de l'Institut international de statistique. La Section on Survey Research Methods de l'American Statistical Association a été établie en 1978, après avoir été une sous-section de la Social Statistics Section entre 1974 et 1977. La Social Statistics Section de la Royal Statistical Society a été fondée en 1976, d'abord sous le nom de Social Statistics and Survey Methodology Study Group. Depuis quelques années, plusieurs de ces associations, l'American Association for Public Opinion Research, ont participé à l'organisation de conférences internationales sur différents aspects des méthodes d'enquête. À noter que plusieurs de ces conférences ont été structurées de façon à

L'évolution de la recherche sur les enquêtes au cours des 25 dernières années

GRAHAM KALTON¹

RÉSUMÉ

Pour marquer le vingt-cinquième anniversaire de *Techniques d'enquête*, l'auteur passe en revue les principales réalisations de la recherche sur les enquêtes au cours des 25 dernières années. Il présente une synthèse générale de l'évolution de la profession, des méthodes d'enquête (conception des questionnaires, collecte des données, traitement des données manquantes, échantillonnage, erreur d'enquête totale) et des applications (enquêtes par panel, enquêtes internationales, analyse secondaire). Il examine enfin les perspectives d'avenir dans ces secteurs.

MOTS CLÉS : Profession de spécialiste des enquêtes; méthodes d'enquête; applications des enquêtes; conception des questionnaires; enquêtes internationales.

1. INTRODUCTION

Techniques d'enquête célèbre cette année son vingt-

cinquième anniversaire. Pour marquer ce jalon, je passerai en revue, dans le présent article, les principales réalisations de la recherche sur les enquêtes au cours des 25 dernières années. Je tiens à souligner, toutefois, que pour plusieurs raisons les dates risquent d'être un peu floues. Tout d'abord, bien sûr, la recherche sur les enquêtes n'a été marquée d'aucun événement spécial en 1975. Au contraire, bon nombre des progrès majeurs survenus au cours du dernier quart de siècle se sont inspirés de travaux antérieurs. Deuxièmement, dans bien des cas, il faut un certain temps avant que les progrès méthodologiques soient acceptés et adoptés intégralement. Troisièmement, mon point de repère est un texte sur la méthodologie des enquêtes que Sir Claus Moser et moi avons publié au Royaume-Uni en 1971 (deuxième édition de *Survey Methods in Social Investigation*, ci-après intitulée *Survey Methods*), la période couverte s'étendant donc en réalité sur plus de 30 ans.

Le présent exposé passe en revue les progrès de la méthodologie des enquêtes, y compris la conception des questionnaires, l'échantillonnage, les méthodes de collecte des données, le traitement des données et l'analyse des enquêtes. L'information est un aspect central, car elle a exercé une influence majeure sur de nombreux progrès méthodologiques, mais pas tous. Le présent exposé examine également l'effet de ces progrès méthodologiques sur les travaux de recherche liés aux enquêtes, y compris l'évolution des enquêtes par panel, des enquêtes internationales et de l'analyse secondaire. J'insisterai surtout sur les enquêtes démographiques, mais je traiterai aussi des enquêtes auprès des établissements. De plus, puisque je me fonde sur mon expérience, l'exposé privilégie sans doute les travaux accomplis aux États-Unis. Avant de me pencher sur les méthodes et la pratique des enquêtes, je compte décrire

d'abord l'énorme accroissement du nombre d'enquêtes et l'établissement explicite de la profession de spécialiste des enquêtes.

2. LA PROFESSION DE SPÉCIALISTE DES ENQUÊTES

L'histoire de la recherche sur les enquêtes relève largement du XX^e siècle. Cette activité a pris son essor dans les années 1930, à connu une croissance considérable au cours de la Seconde Guerre mondiale et a maintenu ensuite un taux de croissance appréciable. En 1975, les enquêtes auprès des ménages et des établissements étaient déjà bien établies comme moyen de répondre aux besoins en données statistiques des décideurs et des chercheurs de nombreux secteurs comme le commerce et la fabrication, l'agriculture, l'emploi et le chômage, les dépenses des familles, la nutrition, la santé, l'éducation, les voyages, le vieillissement et la criminalité. De plus, les enquêtes menées par les universitaires et d'autres chercheurs en sociologie, en science économique, en science politique, en psychologie, en éducation, en travail social et en hygiène publique, quand en études de marché se sont multipliées. Ce secteur a continué de se développer rapidement au cours des 25 dernières années, surtout que les décideurs ont appris à reconnaître la valeur des données d'enquête, tandis que le perfectionnement des méthodes d'enquête a permis aux spécialistes de répondre à la demande de données statistiques. La demande soutenue de données de plus en plus détaillées parmi les décideurs a stimulé le perfectionnement des méthodes d'enquête tout en favorisant l'établissement d'une profession bien implantée de spécialistes des enquêtes. La croissance rapide de la recherche sur les enquêtes est attribuable en partie à un élargissement du choix de thèmes

Dans leur article, Feder, Nathan et Pfeffermann examinent l'échantillonnage répété d'une population à chaque point particulier dans le temps; puis, ils laissent évoluer les effets aléatoires de premier et de deuxième niveaux stochastiquement au cours du temps. Les auteurs examinent en particulier le cas où les unités de deuxième niveau ne font partie de l'échantillon que pendant quelques périodes, ce qui se produit, par exemple, pour de nombreuses enquêtes sur la population active. Ils proposent une méthode d'estimation en deux étapes. Pour commencer, ils ajustent le modèle à deux niveaux indépendamment à chaque point dans le temps pour obtenir les estimations des effets fixes. Puis, ils estiment les paramètres de la série chronologique. Des poids d'échantillonnage peuvent être intégrés à chaque étape pour tenir compte de l'échantillonnage informatif éventuel.

Rivest et Belmonte proposent une estimation conditionnelle de l'erreur quadratique moyenne des estimateurs régionaux subordonnée au modèle de lissage réalisé. Ils proposent un estimateur naturel de cette EQM; cependant, cet estimateur est parfois assez instable si le lissage est important. Ils proposent aussi une correction du biais dans le cas où la distribution des estimateurs directs est asymétrique. Enfin, ils étudient les propriétés de leur estimateur dans les conditions empiriques de Bayes et illustrent leur méthode en se servant des données sur le sous-dénombrement au Recensement du Canada de 1991.

Shao aborde un sujet important, à savoir l'évaluation de l'imputation par les méthodes cold deck. Au fur et à mesure que les progrès en informatique faciliteront le stockage et la consultation de données d'enquête antérieures ou connexes, les méthodes d'imputation fondées sur les données auxiliaires prendront de plus en plus d'importance. Par conséquent, Shao fait un premier pas et compare les résultats de diverses méthodes d'imputation cold deck à ceux d'autres méthodes d'imputation.

Thompson et Frank examinent l'estimation basée sur un modèle dans le cas des plans d'échantillonnage à dépistage de liens. Ce genre de plan d'échantillonnage permet de suivre certains liens d'un répondant à l'autre. L'échantillonnage en réseau et l'échantillonnage en boule de neige sont deux exemples. Après un exposé général du domaine, ils décrivent plusieurs plans d'échantillonnage à dépistage de liens. Puis, ils présentent un modèle graphique des liens qui unissent les membres d'une population. Enfin, ils élaborent, pour ce genre de population, des méthodes d'inférence fondées sur la vraisemblance au moyen de données obtenues par échantillonnage à dépistage de liens.

Thøberge propose de résoudre le problème de poids extrêmes résultant de l'estimateur par calage en relâchant quelque peu les exigences vis-à-vis l'équation de calage. Il s'agit en fait d'un problème de minimisation du même ordre que celui rencontré lors d'une régression avec coefficients de poids («ridge regression»). Il examine aussi d'autres façons de restreindre les poids. Il discute des propriétés asymptotiques des poids calés et donne des conditions nécessaires et suffisantes pour l'existence de poids restreints qui satisfont l'équation de calage. Il discute aussi d'une façon de poser le problème d'estimation en dosant l'importance qu'on accorde à l'équation de calage et donne différentes façons de restreindre les poids qui ne reposent pas sur l'utilisation d'une distance particulière. Finalement, il propose un estimateur avec poids restreints qui est utile pour de petits domaines et aborde les données aberrantes en développant une méthode semblable à celle utilisée pour traiter les poids extrêmes.

Deux notes brèves concluent le présent numéro. Losinger, Garber, Wagner et Hill présentent une étude de cas qui illustre le soin qu'il faut mettre à rajuster les données pour tenir compte de la non-réponse lors de divers cycles d'une enquête. Enfin, Shaffer examine l'estimation des coefficients de régression d'après des données d'enquête si l'on relâche l'hypothèse voulant que les variables auxiliaires soient fixes.

Vous vous souvenez sans doute que le numéro de décembre de *Techniques d'enquête* a été diffusé, à titre expérimental, dans un format électronique au site Web de Statistique Canada. Nous avons également procédé à un sondage sur Internet pour juger de vos réactions à la version électronique de la revue et déterminer vos préférences. Bien que vous ayez manifesté un certain intérêt pour la version électronique, il semble que le moment ne soit pas encore venu de publier régulièrement la revue dans ce format. Nous considérerons certainement de nouveau cette option dans un avenir proche et vos réponses au sondage nous aideront à améliorer toute future version électronique. Entre-temps, nous continuerons de publier une version imprimée de la revue.

Dans ce numéro

Le présent numéro de *Techniques d'enquête* poursuit la commémoration du 25^e anniversaire de la revue marqué par le numéro de décembre 1999. Les sept premiers articles, que des statisticiens de renom spécialisés dans les méthodes d'enquête ont rédigés pour marquer l'occasion, n'ont pu être publiés dans le numéro de décembre faute d'espace. Je tiens à exprimer ici ma gratitude à tous les auteurs qui ont participé à la rédaction de ces deux numéros commémoratifs si spéciaux et mémorables.

Pour lancer ce numéro spécial, Kalton passe en revue les progrès de la recherche sur les enquêtes réalisées ces 25 dernières années, depuis la publication du premier numéro de *Techniques d'enquête*. Il commence par décrire l'évolution de la profession de spécialiste des enquêtes, notamment la création de revues spécialisées et d'associations professionnelles s'adressant aux spécialistes des méthodes d'enquête, ainsi que les aspects internationaux et multidisciplinaires de la profession. Puis, il examine l'évolution des méthodes d'enquête, y compris les méthodes de conception des questionnaires, de collecte des données, d'évaluation de l'erreur non due à l'échantillonnage, d'échantillonnage et d'estimation. Enfin, il souligne l'importance croissante que l'on accorde aux enquêtes par panel et aux enquêtes internationales, ainsi qu'aux sources de données administratives et à l'analyse de données d'enquête.

Bellhousse examine l'évolution parallèle de la réalisation d'enquêtes et de l'informatique au XX^e siècle. Il décrit d'abord l'interaction entre la réalisation de recensements et la mise au point des premières machines à calculer et des premiers ordinateurs numériques. Puis il montre comment les progrès réalisés par la suite dans le domaine du calcul scientifique ont permis d'appliquer des méthodes et des modèles statistiques plus complexes. Pour conclure l'article, il parle de la mise au point de logiciels statistiques pour la réalisation d'enquêtes et de méthodes axées sur des modèles. Bailar examine le rôle de la statistique dans la réalisation des recensements, en insistant particulièrement sur les erreurs de dénombrement dues à des erreurs de recensement de diverses sortes et sur le rajustement des chiffres de recensement d'après l'estimation du sous-dénombrement net fondée sur un échantillon. Elle décrit les diverses sources d'erreurs qui entachent les données de recensement, puis examine les méthodes statistiques d'évaluation des recensements, les méthodes de contrôle de la qualité du traitement des données de recensement et les méthodes d'imputation. Enfin, elle se sert d'un modèle du biais qui entache les données de recensement et de la variance de ces données pour donner une idée de l'efficacité des diverses méthodes de rajustement des données de recensement.

Isaki, Tsay et Fuller étudient l'estimation des facteurs de correction des données de recensement d'après les données de l'enquête postcensitaire de 1990. Leurs estimateurs se fondent sur un modèle des composantes de la variance comprenant un prédicteur linéaire fixe et un effet aléatoire pour décrire le facteur de correction réel inconnu pour chacune des 336 strates définies a posteriori. Ils considèrent d'autres solutions fondées sur l'utilisation d'une estimation de la matrice complète de variances-covariances des erreurs directes d'enquête qui entachent les facteurs de rajustement des strates a posteriori, plutôt que sur l'utilisation des éléments diagonaux uniquement, car cette dernière méthode risque de réduire les effets de l'instabilité de l'estimation de la matrice de variances-covariances complète. Une comparaison empirique les porte à conclure que la meilleure solution est celle qui représente un compromis entre ces deux extrêmes. Ils limitent aussi les facteurs de correction fondés sur le modèle de sorte que l'estimation de la population totale corresponde à celle obtenue d'après les estimations directes d'enquête de ces facteurs de correction.

Lachapelle et Kerr présentent une application innovatrice de l'étude de couverture pour examiner les estimations démographiques. Leur méthode consiste à ventiler les résultats de la contre-vérification des dossiers (CVD) de Statistique Canada afin d'obtenir une source supplémentaire de données que l'on peut comparer aux estimations plus classiques des composantes de la croissance fondées sur les dossiers administratifs. L'objectif de cette comparaison est de repérer les sources principales d'erreur qui entachent les données basées soit sur les dossiers administratifs, soit sur la CVD. Ils montrent aussi que l'on peut décomposer l'erreur de fin de période en deux éléments, à savoir l'écart entre les estimations de la population recensée fondées sur la CVD, d'une part, et sur les données de recensement, d'autre part, et l'écart entre les estimations de la croissance fondées sur la CVD, d'une part, et les dossiers administratifs, d'autre part.

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada
Volume 26, numéro 1, juin 2000

TABLE DES MATIÈRES

Dans ce numéro	1
G. KALTON	
L'évolution de la recherche sur les enquêtes au cours des 25 dernières années	3
D.R. BELLHOUSE	
Évolution de la théorie de l'échantillonnage d'enquête au XX ^e siècle et son rapport avec l'informatique	13
B.A. BAILLAR	
Le passé en guise de prélude	25
C.T. ISAKI, J.H. TSAY et W.A. FULLER	
Estimation des facteurs de correction au recensement	37
R. LACHAPELLE et D. KERR	
Erreur de couverture au recensement: une évaluation démographique	51
M. FEDER, G. NATHAN et D. PFEFFERMANN	
Modélisation multiniveaux des données longitudinales d'enquêtes complexes à effets aléatoires variables en fonction du temps	63
L.-P. RIVEST et E. BELMONTTE	
Une erreur quadratique moyenne conditionnelle des estimateurs régionaux	79
J. SHAO	
Imputation par la méthode cold deck et la méthode du quotient	91
S.K. THOMPSON et O. FRANK	
Estimation fondée sur un modèle et comportant des plans d'échantillonnage à dépistage de liens	99
A. THÉBERGE	
Calage et poids restreints	113
W.C. LOSINGER, L.P. GARBER, B.A. WAGNER et G.W. HILL	
Mise en garde sur la correction des poids selon la non-réponse	123
J. P. SHAFER	
Les meilleurs estimateurs linéaires sans biais locaux et non conditionnels: applications à l'échantillonnage	127

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Techniques d'enquête est répertoriée dans The Survey Statistician, Statistical Theory and Methods Abstracts et SRM Database of Social Research Methodology, Erasmus University. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

COMITÉ DE DIRECTION

Président	G.J. Brackstone
Membres	D. Binder G.J.C. Hole F. Mayda (Directeur de la Production) C. Patrick

COMITÉ DE RÉDACTION

Rédacteur en chef	M.P. Singh, <i>Statistique Canada</i>
Rédacteurs associés	D.R. Bellhouse, <i>University of Western Ontario</i> P. Biemer, <i>Research Triangle Institute</i> D. Binder, <i>Statistique Canada</i> C. Clark, <i>U.S. Bureau of the Census</i> J.-C. Deville, <i>INSEE</i> J. Eltinge, <i>Texas A&M University</i> W.A. Fuller, <i>Iowa State University</i> M.A. Hidroglou, <i>Statistique Canada</i> D. Holt, <i>Central Statistical Office, U.K.</i> G. Kalton, <i>Westat, Inc.</i> P. Kott, <i>National Agricultural Statistics Service</i> S. Lahiri, <i>University of Nebraska-Lincoln</i> S. Limacre, <i>Australian Bureau of Statistics</i> G. Nathan, <i>Central Bureau of Statistics, Israel</i>

Rédacteurs adjoints	J.-F. Beaumont, P. Dick, H. Mantel, B. Quenneville et D. Stukel, <i>Statistique Canada</i>
----------------------------	--

POLITIQUE DE RÉDACTION

Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à faire parvenir le texte rédigé en anglais ou en français au rédacteur en chef, M. M.P. Singh, Division des méthodes d'enquêtes des ménages, Statistique Canada, Tunney's Pasture, Ottawa (Ontario), Canada K1A 0T6. Prière d'envoyer quatre exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

Abonnement

Le prix de Techniques d'enquête (n° 12-001-XPB au catalogue) est de 47 \$ CA par année. Le prix n'inclut pas les taxes de vente canadiennes. Les frais de livraison supplémentaires s'appliquent aux envois à l'extérieur du Canada: États-Unis 12 \$ CA (6 \$ x 2 exemplaires); autres pays, 20 \$ CA (10 \$ x 2 exemplaires). Prière de faire parvenir votre demande d'abonnement à Statistique Canada, Division de la diffusion, Gestion de la circulation, 120, avenue Parkdale, Ottawa (Ontario), Canada K1A 0T6 ou commandez par téléphone au 1 800 700-1033, par télécopieur au 1 800 889-9734 ou par Courriel: order@statcan.ca. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête, l'American Association for Public Opinion Research, la Société Statistique du Canada et l'Association des statisticiennes et statisticiens du Québec.



Ottawa

ISSN 0714-0045

Périodicité: semestrielle

N° 12-001-XPB au catalogue

Juillet 2000

Tous droits réservés. Il est interdit de reproduire ou de transmettre le contenu de la présente publication, sous quelque forme ou par quelque moyen que ce soit, enregistrément sur support magnétique, reproduction électronique, mécanique, photographique, ou autre, ou de l'emmagasiner dans un système de recouvrement, sans l'autorisation écrite préalable des Services de concession des droits de licence, Division du marketing, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

© Ministre de l'Industrie, 2000

Publication autorisée par le ministre
responsable de Statistique Canada

Juin 2000 • VOLUME 26 • NUMÉRO 1

UNE REVUE ÉDITÉE PAR STATISTIQUE CANADA

TECHNIQUES D'ENQUÊTE





NUMÉRO 1

VOLUME 26

JUIN 2000

PAR STATISTIQUE CANADA

UNE REVUE
ÉDITÉE

N^o 12-001-XPB au catalogue



TECHNIQUES D'ENQUÊTE





SURVEY METHODOLOGY



Catalogue No. 12-001-XPB

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

DECEMBER 2000

•

VOLUME 26

•

NUMBER 2





SURVEY METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

DECEMBER 2000 • VOLUME 26 • NUMBER 2

Published by authority of the Minister
responsible for Statistics Canada

© Minister of Industry, 2001

All rights reserved. No part of this publication may be reproduced,
stored in a retrieval system or transmitted in any form or by any
means, electronic, mechanical, photocopying, recording or otherwise
without prior written permission from Licence Services,
Marketing Division, Statistics Canada,
Ottawa, Ontario, Canada K1A 0T6.

February 2001

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa



Statistics
Canada

Statistique
Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is abstracted in The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

MANAGEMENT BOARD

Chairman G.J. Brackstone

Members D.A. Binder
G.J.C. Hole
F. Mayda (Production Manager)
C. Patrick

R. Platek (Past Chairman)
E. Rancourt
D. Roy
M.P. Singh

EDITORIAL BOARD

Editor M.P. Singh, *Statistics Canada*

Associate Editors

D.R. Bellhouse, *University of Western Ontario*

P. Biemer, *Research Triangle Institute*

D.A. Binder, *Statistics Canada*

C. Clark, *U.S. Bureau of the Census*

J.-C. Deville, *INSEE*

J. Eltinge, *Texas A&M University*

W.A. Fuller, *Iowa State University*

J. Gambino, *Statistics Canada*

M.A. Hidirolou, *Statistics Canada*

D. Holt, *Central Statistical Office, U.K.*

G. Kalton, *Westat, Inc.*

P. Kott, *National Agricultural Statistics Service*

P. Lahiri, *University of Nebraska-Lincoln*

S. Linacre, *Australian Bureau of Statistics*

G. Nathan, *Central Bureau of Statistics, Israel*

D. Norris, *Statistics Canada*

D. Pfeffermann, *Hebrew University*

J.N.K. Rao, *Carleton University*

L.-P. Rivest, *Université Laval*

F.J. Scheuren, *The Urban Institute*

R. Sitter, *Simon Fraser University*

C.J. Skinner, *University of Southampton*

E. Stasny, *Ohio State University*

R. Valliant, *Westat, Inc.*

J. Waksberg, *Westat, Inc.*

K.M. Wolter, *National Opinion Research Center*

A. Zaslavsky, *Harvard University*

Assistant Editors J.-F. Beaumont, P. Dick, H. Mantel, W. Yung and D. Stukel, *Statistics Canada*

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their manuscripts in either English or French to the Editor, Dr. M.P. Singh, Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Four nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of Survey Methodology (Catalogue no. 12-001-XPB) is CDN \$47 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN \$12 (\$6 × 2 issues); Other Countries, CDN \$20 (\$10 × 2 issues). Subscription order should be sent to Statistics Canada, Dissemination Division, Circulation Management, 120 Parkdale Avenue, Ottawa, Ontario, Canada K1A 0T6 or by dialling 1 800 700-1033, by fax 1 800 889-9734 or by E-mail: order@statcan.ca. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, the Statistical Society of Canada and l'Association des statisticiennes et staticiens du Québec.

LESLIE KISH
(1910 - 2000)

This issue is dedicated to the memory of Leslie Kish. His infectious joie de vivre, his deep concern for the oppressed and the underprivileged, and his profound contributions to survey methodology and statistics have been, and continue to be, an inspiration to so many.

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Volume 26, Number 2, December 2000

CONTENTS

In This Issue	117
I.P. FELLEGI	
Leslie Kish – A Life of Giving	119
D.E. HAINES, K.H. POLLOCK and S.G. PANTULA	
Population Size and Total Estimation When Sampling From Incomplete List Frames With Heterogeneous Inclusion Probabilities	121
J.-F. BEAUMONT	
An Estimation Method for Nonignorable Nonresponse	131
B.D. SPENCER	
An Approximate Design Effect for Unequal Weighting When Measurements May Correlate With Selection Probabilities	137
P.P. BIEMER and J.M. BUSHERY	
On the Validity of Markov Latent Class Analysis for Estimating Classification Error in Labor Force Data	139
K.J. THOMPSON and R.S. SIGMAN	
Estimation and Replicate Variance Estimation of Median Sales Prices of Sold Houses	153
C.H. McLAREN and D.G. STEEL	
The Impact of Different Rotation Patterns on the Sampling Variance of Seasonally Adjusted and Trend Estimates	163
Y. YOU and J.N.K. RAO	
Hierarchical Bayes Estimation of Small Area Means Using Multi-Level Models	173
F.C. OKAFOR and H. LEE	
Double Sampling for Ratio and Regression Estimation With Sub-sampling the Non-respondents	183
J. PICKERY and G. LOOSVELDT	
Modeling Interviewer Effects in Panel Surveys: An Application	189
M. FUCHS	
Screen Design and Question Order in a CAI Instrument Results From a Usability Field Experiment	199
Acknowledgements	209

In This Issue

This issue is dedicated to Leslie Kish, who passed away this fall at the age of 90. It is remarkable to note that to the end of his life Professor Kish continued to propose and develop new ideas in statistics and survey methodology, as evidenced by his article "Cumulating/Combining Population Surveys" which appeared one year ago in the 25th anniversary issue of this journal. This issue of *Survey Methodology* opens with a reflection on his life and contributions to statistics written by Ivan Fellegi.

The paper by Haines, Pollock and Pantula examines the estimates of a total when two incomplete list frames are combined with an area frame. The authors give suggestions on appropriate population totals to account for the incompleteness of the frames. In addition, their models allow for the fact that larger sampling units are more likely to be included on the incomplete list frames.

Beaumont proposes an estimation method which reduces the bias induced by a response mechanism that depends on the variable of interest, known as a nonignorable response mechanism. The proposed method requires one model for the variable of interest and one model for the response probability. The method is considered robust with respect to the hypothesis of normality since it is constructed in such a way that there is no need to specify the error distribution of model involving the variable of interest, unlike the method of maximum likelihood. The author also proposes a simple method of verifying the validity of the hypothesis of error normality whenever nonresponse is not ignorable.

Spencer considers the problem of estimating the design effect due to weighting when the selection probabilities are correlated with the variable of interest. Using a regression representation of the population, Spencer presents an approximation to the design effect when the selection probabilities are correlated with the variable of interest.

Biemer and Bushery use the Markov assumption on labour force transitions to identify classification errors in labour force data. Using this methodology, they estimate response error rates in panels of monthly labour force data from the Current Population Survey (CPS). The general consistency of the results is taken as an indicator that Markov Latent Class Analysis is a useful method to assess the accuracy of responses in the CPS. Critical to this analysis is confirming the Markov assumption; the authors present some interesting empirical evidence for its validity over the short term in the CPS.

Many statistical offices use modified half-sample-replication (MHS) for estimating the sampling variance of medians. This is an important practical problem because direct calculation of sample medians can be computationally intensive. An alternative estimation method is to group the continuous data into discrete intervals and use linear interpolation over the interval containing the median. In their paper Thompson and Sigman compare the effects of no grouping (*i.e.*, the sample median), grouping with fixed-size intervals, and grouping with data-dependent-sized intervals on medians and associated MHS variance estimates. Their empirical study shows that the data-dependent-sized intervals yielded variance estimates with the smallest bias, the best stability, and the best confidence intervals.

McLaren and Steel consider the implications of different overlap patterns on the sampling variance of seasonally adjusted and trend estimates obtained from time series based on sample surveys by using the Census X-11 and X-11-ARIMA seasonal adjustment methods. They show that the "in for 8", "in for 6", "in for 4, out for 4, in for 4" rotation patterns are sensible if the one month change in seasonally adjusted estimates are the key statistics to be analyzed. If, however, the key statistics are the trend level and the difference between two consecutive trend estimates, then the "in for 1, out for 2, in for 1, for a total of 8 months" is a preferable rotation pattern to reduce the sampling variance. They also show that the "in for 2, out for 2, in for 2, for a total of 8 months" is a reasonable compromise if the level and one months change in seasonally adjusted and trend estimates are both considered important.

You and Rao present hierarchical Bayes multi-level models for small area estimation. The models allow random regression parameters that also depend on small area level covariates. The small area mean is estimated by the posterior mean and the posterior variance is taken as a measure of precision. Three variance models are considered: fixed equal, fixed unequal, and random. Details of Gibbs sampling for these models are presented and used for inference. Procedures are illustrated using county level household income data from Brazil.

Okafor and Lee consider a two phase sampling scenario, where a subsample of the non respondents at the second phase are revisited according to a fixed sampling rate. Based on this scheme, modified versions of the ratio and regression estimators are suggested. Optimal values for the sample sizes and the fixed sampling rate are determined, based on cost functions, so as to minimize variance. In addition variances and their estimators are given. A small empirical study looks at the relative efficiencies of the modified ratio and regression estimators relative to the standard Hansen-Hurwitz estimator.

Pickery and Loosveldt bring an important analytical technique to the study of item non-response. Their models present a more complete picture of the factors affecting item non-response than in previous work in this area. One important aspect of this approach is that the authors make a separation between interviewer/respondent specific variation, variation attributable to interviewer/respondent characteristics and error variance.

Fuchs investigates the affect that screen design and question order have on interviewer behavior in a Computer Assisted Interview (CAI) environment. Through the use of experiments under laboratory conditions, it has been shown that screen design and question order do affect interviewer performance. In his paper, Fuchs presents results from a field experiment which tests two different screen designs together with two different question orders in a 2x2 factor design. These results were based on time measures that were built into the CATI application and from 234 randomly selected interviews that were video taped and analyzed according to a coding scheme.

M.P. Singh

Leslie Kish – A Life of Giving

IVAN P. FELLEGI¹

1. INTRODUCTION

I cannot believe that I am writing an article in memory of Leslie Kish. Just a few months ago I wrote a partly humorous little speech on the occasion of his 90th birthday celebration. I jokingly asked why are we making such a fuss about a 90th birthday – after all the Queen mother just celebrated her 100th. I emphasized that *that* was something. He laughed heartily, with the well known “Kish twinkle” in his eye. I was struck once again by the extent to which he remained fun-loving, vibrant, insightful, in fact *young* in all aspects of behaviour – even if somewhat limited in his mobility. He told me about his forthcoming partial knee replacement operation and confided that his doctor told him that he will either undergo this operation, or he will need to use a walker to get around. Of course, a walker was not to be contemplated: he needed to have his full mobility. And mobility, at 90, meant not just getting around at home but traveling around the world several times a year. He died due post-operative complications, having fought for several weeks with his usual indomitable courage.

In my mind the most characteristic feature of his life was his incessant giving. One of his last acts of giving was to inspire his friends and colleagues to establish the Leslie Kish International Fellows Fund to help students from developing countries obtain training in population sampling.

Leslie was born in 1910 in Poprad, then part of the Austro-Hungarian Empire, now in Slovakia. He used to relate how, at various times throughout history, Poprad belonged to five different countries – an appropriate symbol of his life motivated by a love of people from all parts of the world. In 1925 his parents decided to migrate to the U.S.A – together with hundreds of thousands of other Hungarians who left their country. As the great Hungarian poet Attila Jozsef put it: “one and a half million of our people staggered out to America”. Soon after their arrival Leslie’s father died. The remaining family of mother and four children had to decide whether they will stay in the U.S.A. They did, but that meant that the two oldest children, including Leslie, who was then 16 years old, would have to work in order to help the others.

Leslie continued his schooling in the evening. By 1937 he was within a year of completing his undergraduate studies. But this 27 year old was once again ready to sacrifice himself in order to help the world improve. He interrupted his studies in order help fight the fascists in

Spain as a member of the International Brigade. His love of things Spanish, and of people oppressed, stayed with him forever.

At the end of the Spanish Civil War in 1939 he returned to the United States and completed his studies at City College of New York and received a degree in mathematics. He moved to Washington, where he was fortunate to have become a member of pioneering groups, first at the Bureau of the Census and then at the Department of Agriculture.

Again, he interrupted his career to volunteer for service in the war. In 1947 he finally moved to the University of Michigan at Ann Arbor where, together with a small band of enthusiasts helped found the Institute for Social Research. He said later that he never worked as hard as he did in those early years: obtaining his M.A and Ph.D. while working full time but also finding time to teach.

In statistics, he gave us several superb books. These include the pioneering *Survey Sampling* which became not just a bible of the field (*i.e.*, like the original one, a source of lofty inspiration), as well as a day to day tool of practice. In that sense much of the world’s statistical system has embedded in it the hundreds of pearls of practical wisdom of *Survey Sampling*. In 1988 (when Leslie was a young 78) came *Statistical Design for Research* which integrated and organized a lifetime’s worth of acquired statistical wisdom. In between, before and after came a stream of articles, lectures and talks. He, sometimes working with others, introduced the concepts into our thinking and the words into our language of *design effects*; he was among the first to explore the issue of inference from complex samples and developed the innovation now known as *balanced repeated replication* (actually with Marty Frankel); was among the pioneers of studying *response errors*; became the apostle of *rolling samples and censuses*; pioneered *controlled selection*; formulated the concept of *multipurpose designs*; did some of the early work on *small area estimation*; and so on. But important as these works are, I think just as crucial were some of his other contributions.

He was one of very few people whose early *applied* work made sampling respectable and admired. In addition to having been one of the founders of what became the *Institute for Survey Research* at Ann Arbor, he taught generations of statisticians, both Americans and foreign ones through the legendary Summer Program for Foreign Statisticians. After his formal retirement he continued to do so through lectures in the Summer Program; through

¹ Ivan P. Fellegi, Chief Statistician, Statistics Canada, 26th floor, section A, R.-H. Coats Building, Ottawa, Ontario, Canada K1A 0T6.

decades of editing or contributing to one or another of the questions and answers columns of the *Survey Statistician*; and through numerous lectures and consulting assignments. At international meetings I used to “bump into” his past students and current friends. One no longer “bumps into” them, because they have become completely ubiquitous: I wonder how many better known foreign samplers there are who were *not* at some point Leslie’s students. And I do not want to forget about two of my favourites among his many contributions. His years of faithful service to Statistics Canada as a founding member of our Advisory Committee on Statistical Methods; and his ASA presidential address of 1977 (published in *JASA* in March 1978) – the best address that any President of ASA gave in my living memory.

For his accomplishments he received world wide recognition. Of his dozens of awards I will just single out a few: he received an honorary doctorate from the University of Bologna on the occasion of its 900th anniversary, the Samuel Wilks Medal which is ASA’s highest recognition, the Henry Russell lectureship which is the highest recognition of University of Michigan, the title Honorary Fellow of the ISI which I regard as a kind of Nobel prize in statistics, and perhaps the most personally meaningful for

him: a slew of the highest possible recognitions from Hungary (honorary doctorate from the largest university in Budapest, honorary membership in the Hungarian Academy of Sciences and the Officer’s Cross of the Order of the Merit).

Over and above what he gave us in statistics, he gave us the phenomenon known as “Leslie Kish, a force of nature”: the Spanish Civil War fighter, the philosopher of all things statistical, the ever young agitator for human rights, raconteur, avid reader, author of the best annual Christmas letters, loving husband and father, and lifelong friend to hundreds, perhaps thousands.

When I spoke at his 90th birthday celebration, I ended by saying that I was hoping to be present at Leslie’s really big anniversary – the one the Queen Mother had just passed. And that was not just a joke: he was so full of life, it was not only quite possible to contemplate him living to be a hundred, but rather it was impossible to think about the opposite. Unfortunately, he did pass away. His final act of giving was to donate his body to medical research. Wouldn’t it be fitting if the resulting work gave us some insight into the human wonder that was Leslie Kish?...

Population Size and Total Estimation When Sampling From Incomplete List Frames With Heterogeneous Inclusion Probabilities

DAWN E. HAINES, KENNETH H. POLLOCK and SASTRY G. PANTULA¹

ABSTRACT

Information from list and area sampling frames is combined to obtain efficient estimates of population size and totals. We consider the case where the probabilities of inclusion on the list frames are heterogeneous and are modeled as a function of covariates. We adapt and modify the methodology of Huggins (1989) and Alho (1990) for modeling auxiliary variables in capture-recapture studies using a logistic regression model. We present the results from a simulation study which compares various estimators of frame size and population totals using the logistic regression approach to modeling heterogeneous inclusion probabilities.

KEY WORDS: Logistic regression; List frame; Area frame; Capture-recapture sampling.

1. INTRODUCTION

In this paper, we estimate population size and totals when information from multiple independent sampling frames is available. Population elements are assumed to have varying probabilities of inclusion for different sampling frames. These heterogeneous inclusion probabilities may depend on a covariate. For example, suppose we are interested in estimating the number of hog farms and the total number of hogs in North Carolina. Covariate measurements such as hog farm acreage or number of employees indicate the size of hog farms. Larger farms may have a higher chance of being included on a list frame than smaller farms. In capture-recapture experiments, animals may have unequal capture probabilities. Capture (inclusion) probabilities for animals may vary with respect to age, sex, size, or species.

List frames are physical listings of sampling units in the target population. Items found on a list frame can include, but are not limited to, names, addresses, telephone numbers, social security numbers, or physical descriptions of locations. These and other miscellaneous stratification variables are used to identify persons, animals, businesses, or other establishments. List and area sampling frames are constructed and maintained to obtain estimates of the unknown population size and totals. Since frame imperfections such as omissions, duplications, and inaccurate recordings are inevitable in any large data collection operation (Hansen, Hurwitz and Madow 1953), various solutions for dealing with frame imperfections have been proposed in the literature. One approach, first developed by Hartley (1962, 1974), combines an incomplete list frame with an area frame. Further theoretical extensions are due to Cochran (1965), Lund (1968), Fuller and Burmeister (1972), and Bosecker and Ford (1976). Haines and Pollock (1998a) apply the dual frame method to a bald eagle population

while Haines and Pollock (1998b) present a more general, theoretical approach to combining multiple frames. These two papers do not consider the case where the inclusion probabilities are heterogeneous. Fienberg (1992) presents an annotated bibliography of the capture-recapture literature specifically related to the census undercount problem, including Wolter (1986, 1990), and Cowan and Malec (1986).

The National Agricultural Statistics Service (NASS) currently employs a multi-frame approach for its sampling and estimation of numerous agricultural commodities. NASS collects and summarizes data on crop acreage, livestock, grain production and stocks, costs of production, farm expenditures, and other agricultural items. Fecso, Tortora and Vogel (1986) provide a review of sampling frames for the agricultural sector of the United States while Nealon (1984) details the multiple and area frame estimators used by the U.S. Department of Agriculture. Pollock, Turner and Brown (1994) offer a model-based capture-recapture solution for estimating frame size based on information from two incomplete list frames. According to Cochran (1977), it is often difficult to obtain a list that corresponds exactly to the population of interest. Lists routinely collected for some purpose are usually found to be incomplete, partially illegible, or to contain an unknown amount of duplication. Since list frames are typically incomplete, estimates based solely on list frames may underestimate the population size. Supplementing available information with an area frame sample may provide efficient estimates of the population size and totals.

An area frame is a collection of geographical areas defined by identifiable boundaries. Area frames are often used by survey practitioners in order to attain complete coverage of the target population. Populations such as farms are naturally associated with the land units comprising the area frame. For example, in an agricultural survey, the region of

¹ Dawn E. Haines, U.S. Bureau of the Census, Washington, DC 20233; Kenneth H. Pollock and Sastry G. Pantula, North Carolina State University, Department of Statistics, Box 8203, Raleigh, NC 27695-8203, U.S.A.

interest is divided into a set of geographic land masses called segments. Segments, which are the sampling units, are then selected using stratified multistage designs (Kott and Vogel 1995). Rules which link farms in the population to segments in the area frame are defined. Once the farms, or reporting units, within each sampled segment are identified, they are personally enumerated and the pertinent data collected. Nealon (1984) provides a detailed description of the open, closed, and weighted segment estimators. Faulkenberry and Garoui (1991) formulate additional estimators specifically designed for area frames. More complex construction and sampling methods for area frames are discussed in Fecso *et al.* (1986). Area sampling and subsampling from area frames are considered in detail in Kott and Vogel (1995).

In section 2, we consider independent list frames where the list frame elements have heteroscedastic inclusion probabilities. We discuss methods which provide population size and total estimators when information from list frame(s) and an area frame sample is available. Section 3 summarizes results from a simulation study that compares various estimators of frame (population) size and population totals. Finally, results are summarized and discussed.

2. HETEROSCEDASTIC INCLUSION PROBABILITIES

2.1 Population Size Estimation with List Frames

In capture-recapture experiments, different animals may have different capture probabilities. Similarly, individual elements may have different probabilities of inclusion on a list frame. Different list frames may be viewed as different capture occasions. Model M_h denotes the heterogeneity model in the closed population capture-recapture literature (Otis, Burnham, White and Anderson 1978). In a capture-recapture setting, capture probabilities, though assumed to vary from animal to animal, are assumed to be the same for all trapping occasions. The heterogeneity model may have up to $N + 1$ total parameters, namely N and p_i , $i = 1, \dots, N$, where N is the population size and p_i denotes the inclusion probability for the i -th unit. For multiple list frames, this corresponds to the assumption that the inclusion probability p_i for element i is constant over all k list frames, B_1, B_2, \dots, B_k .

Burnham (1972) and Burnham and Overton (1978, 1979) investigate the problem of estimating N in the capture-recapture setting. The proposed estimator for N given by Burnham (1972) is based on the jackknife method of bias reduction (Quenouille 1956). Chao (1988) develops an alternative moment estimator for this model based on capture frequency data (Pollock 1991). Under certain conditions, Chao's proposed estimator is less biased than Burnham's jackknife estimator. In general, it is difficult to find a completely satisfactory estimator of N under Model M_h . Otis *et al.* (1978), as a result, suggest that one should design

the entire study to minimize heterogeneity. Norris and Pollock (1996) propose a nonparametric MLE which is still not totally satisfactory.

In capture-recapture experiments, the model expressed as Model M -th allows inclusion probabilities to vary both by trapping occasion (list frame) and individual. Define p_{ij} as the inclusion probability of the i -th element on the j -th list frame. Model M -th is obviously not easy to estimate since it can have up to $tN + 1$ parameters where $t = k$, the number of list frames. Chao, Lee and Jeng (1992), using the idea of sample coverage, propose a nonparametric method of estimating the population size for Model M -th.

An alternative to the nonparametric approach is to model the inclusion probabilities as a function of an auxiliary variable. Pollock, Hines and Nichols (1984), Huggins (1989), and Alho (1990) address the role of auxiliary variables in capture-recapture experiments with unequal capture (inclusion) probabilities. The closed population capture-recapture experiments have $i = 1, \dots, N$ individuals and $j = 1, \dots, t$ trapping occasions. Again, the $j = 1, \dots, t$ trapping occasions are similar to $t = k$, the number of independent list frames. Huggins (1989) and Alho (1990) propose a conditional estimation procedure for estimating the size of a closed population based on one capture and a single recapture. Both of these papers assume the logistic model for the inclusion probabilities, given by

$$p_{ij} = \frac{\exp(\alpha_j + \beta_j x_i)}{1 + \exp(\alpha_j + \beta_j x_i)}, \quad (1)$$

where x_i is a covariate α_j and β_j are unknown parameters. Note that this parameterization yields $0 \leq p_{ij} \leq 1$ for all values of α_j and β_j . For $\beta_j > 0$, the inclusion probability increases with the covariate. This parameterization is different from the probability proportional to size (pps) sampling where p_{ij} is assumed to be proportional to x_i . The MLEs of α_j and β_j can be obtained using the likelihood conditioned on the unit being on at least one list frame. Haines (1997) derives the conditional likelihood function for three independent list frames.

Treating each individual as a separate stratum, define the following indicator variables for $i = 1, \dots, N$:

$$u_{ij} = \begin{cases} 1 & \text{individual } i \text{ belongs to frame } j \text{ only} \\ 0 & \text{otherwise} \end{cases}$$

$$j = B_1, B_2,$$

and

$$a_i = \begin{cases} 1 & \text{individual } i \text{ belongs to both frames} \\ 0 & \text{otherwise.} \end{cases}$$

The value of the expression

$$M_i = u_{iB_1} + u_{iB_2} + a_i \quad (2)$$

is one if individual i is included on at least one of the two frames and zero otherwise.

Alho (1990) presents the conditional likelihood function for two list frames as

$$\frac{\exp\{\alpha_{B_1} N_{B_1} + \alpha_{B_2} N_{B_2} + \beta_{B_1} \sum_{i \in B_1} x_i + \beta_{B_2} \sum_{i \in B_2} x_i\}}{\prod_{M_i=1} K_i(\theta)}, \quad (3)$$

where

$$K_i(\theta) = \exp\{\alpha_{B_1} + \beta_{B_1} x_i\} + \exp\{\alpha_{B_2} + \beta_{B_2} x_i\} \\ + \exp\{\alpha_{B_1} + \alpha_{B_2} + (\beta_{B_1} + \beta_{B_2}) x_i\}$$

and $\theta = (\alpha_{B_1}, \beta_{B_1}, \alpha_{B_2}, \beta_{B_2})'$. Alho (1990) uses an iterative procedure based on the sufficient statistics to maximize (3) while we implement Newton's method to calculate conditional MLEs of θ , denoted $\hat{\theta} = (\hat{\alpha}_{B_1}, \hat{\beta}_{B_1}, \hat{\alpha}_{B_2}, \hat{\beta}_{B_2})'$. See Appendix A of Haines (1997) for details on Newton's method. The estimated probability that individual i is included on at least one list frame is denoted

$$\hat{\pi}_i = 1 - \left(\frac{1}{1 + \exp(\hat{\alpha}_{B_1} + \hat{\beta}_{B_1} x_i)} \right) \\ \times \left(\frac{1}{1 + \exp(\hat{\alpha}_{B_2} + \hat{\beta}_{B_2} x_i)} \right) = \pi_i(\hat{\theta}), \quad (4)$$

where

$$\hat{p}_{ij} = \frac{\exp(\hat{\alpha}_j + \hat{\beta}_j x_i)}{1 + \exp(\hat{\alpha}_j + \hat{\beta}_j x_i)}, \\ i = 1, \dots, N \text{ and } j = B_1, B_2. \quad (5)$$

If θ were known, the Horvitz-Thompson estimator of N is $\hat{N} = \sum_{M_i=1} 1/\pi_i$ (Horvitz and Thompson 1952). From Cochran (1977), the variance of \hat{N} is

$$V(\hat{N}) = \sum_{i=1}^N \frac{1 - \pi_i}{\pi_i}. \quad (6)$$

An estimate of the variance of \hat{N} is

$$\hat{V}(\hat{N}) = \sum_{M_i=1} \frac{1 - \pi_i}{\pi_i^2}.$$

Since θ is unknown, we consider the population size estimate given by $\hat{N} = \sum_{M_i=1} 1/\hat{\pi}_i$, where $\hat{\pi}_i$ is defined in (4). An estimate of the variance of \hat{N} is derived using Taylor's method and has the form

$$\hat{V}(\hat{N}) = \sum_{M_i=1} \frac{1 - \pi_i(\hat{\theta})}{\pi_i^2(\hat{\theta})} + \hat{A} \sum (\hat{\theta}) \hat{A}', \quad (7)$$

where

$$\hat{A} = \sum_{M_i=1} \left[\frac{1}{\pi_i^2(\hat{\theta})} \frac{\partial \pi_i(\hat{\theta})}{\partial \hat{\theta}'} \right]$$

and $\sum(\hat{\theta})$ is the inverse of the Hessian matrix. The second term in (7) is due to estimating $\pi_i(\theta)$ by $\pi_i(\hat{\theta})$.

Another population size estimator commonly used in capture-recapture experiments is the Lincoln-Petersen estimator. This classic estimator is due to Lincoln (1930) and Petersen (1896) and has the form

$$\hat{N}_{L-P} = \frac{N_{B_1} N_{B_2}}{N_{b_1 b_2}},$$

where N_{B_1} and N_{B_2} denote the size of list frames B_1 and B_2 , respectively, and $N_{b_1 b_2}$ denotes the number of units common to both frames. This is a simple method of moments estimator based on the assumption that all units have homogeneous inclusion probabilities for each of the two independent list frames. It is possible for the denominator $N_{b_1 b_2}$ to be zero. Chapman (1951) proposed a modified version of the Lincoln-Petersen estimator, given by

$$\hat{N}_{CH} = \frac{(N_{B_1} + 1)(N_{B_2} + 1)}{(N_{b_1 b_2} + 1)} - 1. \quad (8)$$

This estimator is less biased than the Lincoln-Petersen estimator (Chapman 1951). According to Sekar and Deming (1949), the asymptotic standard error of \hat{N}_{CH} is

$$\sqrt{\hat{V}(\hat{N}_{CH})} = \sqrt{\frac{N_{B_1} N_{B_2} N_{b_1} N_{b_2}}{(N_{b_1 b_2})^3}},$$

where N_{b_1} and N_{b_2} denote the number of units belonging only to list frames B_1 and B_2 , respectively.

The Lincoln-Petersen estimator is the unconditional maximum likelihood estimator of the population size N when there are two independent list frames and the inclusion probabilities are homogeneous. Haines (1997) extends the estimation procedures to k list frames, each with homogeneous inclusion probabilities. This estimator, however, is not appropriate when the inclusion probabilities are heterogeneous. See the simulation results in section 3.

2.2 Population Size Estimation with Area and List Frames

Suppose we have access to an area frame in addition to two list frames, B_1 and B_2 . The area frame consists of U_A segments that cover the entire population. A simple random sample of u_A segments is selected. We assume that all units in the sampled segments are observed. The probability of inclusion in the area frame sample is the same for all units and is the known quantity $p_A = u_A/U_A$. Next, we maximize the conditional likelihood (3) with respect to θ and calculate the estimated probability that individual i is included on at least one list frame or the area frame. This probability is denoted $\hat{\pi}_i = \hat{\pi}_i + p_A(1 - \hat{\pi}_i)$. The probabilities $\hat{\pi}_i$ and \hat{p}_{ij} are defined in (4) and (5), respectively. An estimated Horvitz-Thompson estimator for population size is

$$\hat{N} = \sum_{i \in \text{sample}} \frac{1}{\hat{\pi}_i} \quad (9)$$

This estimator can easily be extended to the case with k list frames, B_1, \dots, B_k , and an independent area frame.

From Cochran (1977), an estimate of the variance of \hat{N} is given by

$$\hat{V}(\hat{N}) = \sum_{M_i=1} \frac{1 - \hat{\pi}_i}{\hat{\pi}_i^2} + 2 \sum_i \sum_{l < i} \frac{(\hat{\pi}_{il} - \hat{\pi}_i \hat{\pi}_l)}{\hat{\pi}_{il} \hat{\pi}_i \hat{\pi}_l} + \hat{A} \hat{\Sigma} \hat{A}', \quad (10)$$

where \hat{A} is defined in (7) and $\hat{\Sigma}$ is the inverse of the Hessian matrix. The variance formula for \hat{N} in (6) and its estimate are valid only when π_{il} , the probability that units i and l are included in the sample, is equal to $\pi_i \pi_l$. When an area frame sample is included, however, π_{il} is not necessarily equal to $\pi_i \pi_l$. Suppose units i and l belong to the same area frame segment. In this case, units i and l are both included or not included in the sample, depending on whether their corresponding segment is selected or not. It can be shown that the joint inclusion probability, π_{il} , can be estimated as

$$\hat{\pi}_{il} = \begin{cases} \hat{\pi}_i \hat{\pi}_l & \text{if } i \text{ and } l \text{ belong to different} \\ & \text{area segments} \\ p_A + \hat{\pi}_i \hat{\pi}_l (1 - p_A) & \text{if } i \text{ and } l \text{ belong to the same} \\ & \text{area segment} \end{cases} \quad (11)$$

where $\hat{\pi}_i$ is defined in (4) and $\hat{\pi}_i = \hat{\pi}_i + p_A(1 - \hat{\pi}_i)$. Hence, when i and l belong to the same segment, $\hat{\pi}_{il} \neq \hat{\pi}_i \hat{\pi}_l$. However, if p_A is small and $\hat{\pi}_i$ and $\hat{\pi}_l$ are large, then $(\hat{\pi}_{il} - \hat{\pi}_i \hat{\pi}_l)$ will be close to zero.

2.3 Population Total Estimation with List Frames

Suppose y_i values are available for all elements on two independent list frames B_1 and B_2 . If θ were known, an estimate of the population total, Y , is the Horvitz-Thompson estimator

$$\hat{Y}_{H-T} = \sum_{M_i=1} \frac{y_i}{\pi_i(\theta)} \quad (12)$$

According to Cochran (1977), the estimated variance of \hat{Y}_{H-T} is

$$\hat{V}(\hat{Y}_{H-T}) = \sum_{M_i=1} \frac{y_i^2(1 - \pi_i(\theta))}{\pi_i^2(\theta)}$$

When θ is unknown and is estimated by $\hat{\theta}$, an estimate for the population total is

$$\hat{\hat{Y}}_{H-T} = \sum_{M_i=1} \frac{y_i}{\pi_i(\hat{\theta})}$$

An estimate of the variance of $\hat{\hat{Y}}_{H-T}$ is derived using Taylor's method and has the form

$$\hat{V}(\hat{\hat{Y}}_{H-T}) = \sum_{M_i=1} \frac{y_i^2(1 - \pi_i(\hat{\theta}))}{\pi_i^2(\hat{\theta})} + \hat{B} \hat{\Sigma}(\hat{\theta}) \hat{B}',$$

where

$$\hat{B} = \sum_{M_i=1} \left[\frac{y_i}{\pi_i^2(\hat{\theta})} \frac{\partial \pi_i(\hat{\theta})}{\partial \hat{\theta}'} \right]$$

and $\hat{\Sigma}(\hat{\theta})$ is the inverse of the Hessian matrix evaluated at $\hat{\theta}$. These ideas extend easily to incorporate k independent list frames.

In practice, y_i 's may not be observed for all units on the list frames. Consider the case where y_i 's are available for only a random sample of n_{B_1} and n_{B_2} units from list frames B_1 and B_2 , respectively. By construction, the inclusion probabilities, p_{ij} , vary with the individual i and frame j . However, once individuals are included on a list frame, they are subsampled using simple random sampling. As a result, all elements on list frame B_j have equal chance (n_{B_j}/N_{B_j}) of inclusion in the subsample. Note that we are selecting samples from each list frame rather than drawing a single sample from a combined list frame. Since the list frames are assumed to be independent, the estimated probability the i -th individual is included on at least one of the two list frames is

$$\hat{\pi}_i = \hat{p}_{iB_1} \frac{n_{B_1}}{N_{B_1}} + \hat{p}_{iB_2} \frac{n_{B_2}}{N_{B_2}} - \hat{p}_{iB_1} \hat{p}_{iB_2} \frac{n_{B_1}}{N_{B_1}} \frac{n_{B_2}}{N_{B_2}} \quad (14)$$

An estimated Horvitz-Thompson estimate of Y is obtained by substituting (14) into (12).

Another estimator of the population total, Y , in this case is

$$\hat{Y} = \hat{N}_{CH} \frac{\sum_{M_i=1} y_i}{N_{b_1} + N_{b_2} + N_{b_1 b_2}};$$

which is Chapman's estimator multiplied by the mean of the responses for those elements included on at least one list frame subsample. Again, this estimator is valid only when the inclusion probabilities are homogeneous. There are $N_{b_1} + N_{b_2} + N_{b_1 b_2}$ unique elements in frames B_1 and B_2 . A similar estimator can be defined when information is available only for subsamples from the list frames.

2.4 Population Total Estimation with Area and List Frames

Consider the case where, in addition to y_i values for the units on the list frames (or subsamples from list frames), y_i values are available for all elements in a random sample of segments from an area frame. Inclusion of the area frame information results in the estimated inclusion probability for the i -th individual, namely

$$\tilde{\pi}_i = \hat{\pi}_i + p_A(1 - \hat{\pi}_i), \quad (15)$$

where $\hat{\pi}_i$ is defined in (4) or (14), depending on whether y_i is observed for all units on the list frames or only for a subsample of units, respectively. An estimated Horvitz-Thompson estimator of the population total in this case is

$$\hat{Y}_{H-T} = \sum_{i \in \text{sample}} \frac{y_i}{\tilde{\pi}_i}. \quad (16)$$

An estimate of the variance of \hat{Y}_{H-T} is given by

$$\begin{aligned} \hat{V}(\hat{Y}_{H-T}) &= \sum_{M_i=1} \frac{y_i^2(1 - \tilde{\pi}_i)}{\tilde{\pi}_i^2} \\ &+ 2 \sum_{i < l} \sum \frac{(\tilde{\pi}_{il} - \tilde{\pi}_i \tilde{\pi}_l)}{\tilde{\pi}_{il} \tilde{\pi}_i \tilde{\pi}_l} y_i y_l + \hat{B} \hat{\Sigma} \hat{B}', \quad (17) \end{aligned}$$

where $\tilde{\pi}_{il}$ is defined in (11) and \hat{B} and $\hat{\Sigma}$ are defined in (13).

3. SIMULATION STUDY

3.1 Assumptions of the Study

To study the properties of population size and total estimators, Haines (1997) considered eighty different models. Details for only two of those models are presented here. One assumption made is that the inclusion probabilities for two list frames depend on a covariate x_i . Secondly, we assume that the covariate may be correlated with the response variable y_i . Also, we assume that x_i and y_i are lognormally distributed with correlation ρ_{xy} . The lognormal distribution is utilized which allows for a skewed distribution of covariates. We generate x_i as e^{u_i} and y_i as e^{v_i} , where u_i and v_i are generated as bivariate normal random variables with zero means, unit variances, and correlation ρ_{uv} . It can be shown that $\rho_{uv} = \log[\rho_{xy}(e-1) + 1]$.

Consider a population of size N . Assume that there are two independent list frames, B_1 and B_2 , and an area frame, A . The area frame is assumed to be complete in the sense that it covers the entire population. A sample of area frame segments is selected and the units within each area segment are observed. Let p_A denote the inclusion probability for any element to be included in the area frame sample, where p_A is assumed to be the same for all individuals.

The probability that the i -th element is included on the j -th list frame is given by the logistic regression model (1) for $i = 1, \dots, N$ and $j = B_1, B_2$. We assume the probability that the i -th element is included on list frame B_1 is independent of its inclusion status on list frame B_2 and the area frame sample.

3.2 Parameter Settings

We consider various parameter values. For the population size, N , we take $N = 300$ or $1,000$. We use $\rho_{xy} = -0.3, 0.0, 0.5$, and 1 corresponding to negative, zero, positive, and perfect correlation between the response variable and the covariate. Here, $\rho_{xy} = 1$ corresponds to $x_i = y_i$, indicating that the inclusion probability is directly related to the response variable.

For each of the above $2 \times 4 = 8$ parameter settings of N and ρ_{xy} , we consider two models corresponding to different choices of $\alpha_{B_1}, \beta_{B_1}, \alpha_{B_2}$, and β_{B_2} . Recall that $E(x_i) = E[e^{u_i}] = e^{0.5}$. Consider an element with covariate value given by the mean value $e^{0.5}$. The probability that this element is included on the j -th list frame is

$$p_j^{(E)} = \frac{\exp(\alpha_j + \beta_j e^{0.5})}{1 + \exp(\alpha_j + \beta_j e^{0.5})}, \quad j = B_1, B_2.$$

If $\alpha_j = -\beta_j e^{0.5}$, then this element has a 50% chance of being included on list frame j . We use this relationship in Model 1.

Extending the above idea, if we set

$$\alpha_j = \log\left(\frac{p}{1-p}\right) - \beta_j e^{0.5},$$

then the unit with mean covariate value has probability p of being included on list frame j . If we assume that the inclusion probabilities are the same for list frames B_1 and B_2 , then the chance of being included on at least one of the two list frames is given by $1 - (1-p)^2$. This relationship is used in Model 2. Specific values of α_j and β_j for the two models are summarized in Table 1.

Table 1
Summary of Model Parameters

Model	α_{B_1}	β_{B_1}	α_{B_2}	β_{B_2}	$p_{B_1}^{(E)}$	$p_{B_2}^{(E)}$	$1 - (1 - p_{B_1}^{(E)})(1 - p_{B_2}^{(E)})$
1	0	0	0	0	0.5	0.5	0.75
2	-0.5478	0.8	-0.5478	0.8	0.6838	0.6838	0.90

For each of the $2 \times 4 \times 2 = 16$ models, we consider three p_A values given by 0, 0.05, and 0.20. Here, $p_A = 0$ corresponds to using only the information from list frames B_1 and B_2 .

3.3 Generation of the Data

For each of the above sixteen models, we first generate (x_i, y_i) using the bivariate lognormal distribution for $i = 1, \dots, N$. We then "generate" (identify) the units that belong to list frames B_1 and B_2 . We use the probability p_{ij} to include the i -th element on list frame j . Finally, using $p_A = 0.05$, we identify the elements belonging to area frame A . We repeat the process for the case $p_A = 0.20$. For each parametric combination, we generate 1,000 Monte Carlo replications.

3.4 Estimators

For population size, we consider Chapman's estimator, \hat{N}_{CH} , given in (8). This estimator assumes that $\beta_{B_1} = \beta_{B_2} = 0$ and does not utilize the information from the area frame sample. We also consider the estimated Horvitz-Thompson estimators discussed in section 2.

For estimating the population total of a response variable, we consider the case where the response is observed for all list frame elements. Elements in an area frame are sampled with probabilities $p_A = 0, 0.05$, and 0.20 . We do not consider population total estimates based on subsamples from each list frame. The population total estimate, \hat{Y}_{p_A} , has the same form as (16) with $\hat{\pi}_i$ defined in (15). Similarly, the population size estimate, \hat{N}_{p_A} , has the same form as (9).

The estimator

$$\hat{Y}_{CHs, p_A} = \hat{N}_{CH} \bar{y}_{(p_A)}, \quad p_A = 0, 0.05, 0.20$$

is also considered where $\bar{y}_{(p_A)}$ is the sample mean of the y_i 's included in the "sample." The performance of \hat{Y}_{CHs, p_A} is dependent on \hat{N}_{CH} , which was observed to underestimate N considerably for Model 2. The results for \hat{Y}_{CHs, p_A} are not included here. Another design-unbiased estimator of Y is given by

$$\hat{Y}_A = \sum_{i \in \text{"area sample"}} \frac{y_i}{p_A}.$$

This is the Horvitz-Thompson estimator based on the area frame sample alone. Since complete enumeration of area segments is expensive, p_A is typically small in practice. For small p_A , \hat{Y}_A is expected to have a much larger variance than \hat{Y}_{p_A} since the estimator \hat{Y}_{p_A} includes information from list frames in addition to information from the area frame samples. Hence, results for \hat{Y}_A are not included.

3.5 Estimated Variance of the Estimator

In our simulation study the values of p_A considered are very small. In contrast, the probability of inclusion on at least one of the list frames is large for each individual. As a result, $\hat{\pi}_i$ is close to π_i and $\hat{\pi}_{il}$ in (11) is close to $\pi_i \pi_l$. Hence, the second term in equations (10) and (17), involving $\hat{\pi}_{il} - \pi_i \pi_l$, are expected to be small. We have not included this term in our estimate of the variance. Despite this omission, we observe that the estimated variance is very close to the empirical variance of the estimator for the models we consider.

3.6 Summary Statistics

For the population size estimates, we present results averaged over the 4,000 replications corresponding to the four values of p_{xy} and 1,000 Monte Carlo replications for each p_{xy} . For each model, we summarize the mean and standard deviation of the estimates, average of the estimated standard errors of the estimators, the percent relative root

mean square error (% RRMSE), and the empirical coverage probabilities of a 95% confidence interval. These measures are all standardized by the population size N . We report results for Models 1 and 2 in Tables 2 and 3, respectively.

Table 2
Population Size Estimates for Model 1

$N = 300$	\hat{N}_{CH}	\hat{N}_0	$\hat{N}_{0.05}$	$\hat{N}_{0.20}$
Average of estimates divided by N	0.999	1.011	1.007	1.004
Standard deviation of estimates divided by N	0.059	0.077	0.059	0.048
Average of estimated standard deviation of estimator divided by N	0.059	0.072	0.059	0.047
% RRMSE	0.003	0.006	0.004	0.002
Coverage	0.947	0.955	0.957	0.950
$N = 1,000$				
Average of estimates divided by N	1.000	1.003	1.002	1.002
Standard deviation of estimates divided by N	0.031	0.035	0.030	0.025
Average of estimated standard deviation of estimator divided by N	0.032	0.034	0.030	0.025
% RRMSE	0.001	0.001	0.001	0.001
Coverage	0.954	0.959	0.958	0.956

Table 3
Population Size Estimates for Model 2

$N = 300$	\hat{N}_{CH}	\hat{N}_0	$\hat{N}_{0.05}$	$\hat{N}_{0.20}$
Average of estimates divided by N	0.922	1.006	1.005	1.003
Standard deviation of estimates divided by N	0.032	0.052	0.049	0.040
Average of estimated standard deviation of estimator divided by N	0.028	0.052	0.048	0.040
% RRMSE	0.007	0.003	0.002	0.002
Coverage	0.271	0.953	0.954	0.951
$N = 1,000$				
Average of estimates divided by N	0.921	1.001	1.001	1.001
Standard deviation of estimates divided by N	0.018	0.028	0.027	0.022
Average of estimated standard deviation of estimator divided by N	0.015	0.027	0.026	0.021
% RRMSE	0.007	0.0008	0.0007	0.0005
Coverage	0.009	0.949	0.949	0.949

Similarly, for the population total estimates, we present summary statistics averaged over the 1,000 replications corresponding to each parametric combination. We summarize the mean and standard deviation of the estimates as well as the average of the estimated standard errors of the estimators, where the estimates are scaled by the true total (Y) for that replicate. In other words, for each replicate we divide the estimate by its replicate total, Y . We then compute the mean and the standard deviations of these standardized estimates. Similarly, for each replicate, we compute the estimated standard error of the total estimator divided by the total for the replicate and then compute the average of these standardized values. We report these because the totals change from replicate to replicate.

Finally, we report the coverage probabilities of the 95% confidence intervals for the total. Results for Models 1 and 2 are respectively presented in Tables 4 and 5.

3.7 Conclusions

3.7.1 Population Size Estimation

In Model 1, the inclusion probabilities do not depend on the covariate. In this case, Chapman's estimator \hat{N}_{CH} is very close to the maximum likelihood (Lincoln-Petersen) estimator and hence is expected to perform better than \hat{N}_0 .

The estimator \hat{N}_0 loses efficiency since it estimates the parameters α_{B_1} , β_{B_1} , α_{B_2} , and β_{B_2} , which have the value zero in this model. The estimator $\hat{N}_{0.05}$ has about the same efficiency as \hat{N}_{CH} . The bias in all the estimates is minimal. For Model 1, we notice that the average of the estimated standard deviation is close to the standard deviation of the estimates. This indicates that the standard error estimate we use performs well. Also, we notice that the empirical coverage probabilities are all within three standard errors of 0.95. That is, all of the empirical coverage probabilities are within $(0.95 \pm 3 [0.95 \times 0.05/4,000]^{1/2}) = (0.94, 0.96)$.

Table 4
Population Total Subsampling Estimates Scaled by Y for Model 1

		$N = 300$			$N = 1,000$		
ρ_{xy}		\hat{Y}_0/Y	$\hat{Y}_{0.05}/Y$	$\hat{Y}_{0.20}/Y$	\hat{Y}_0/Y	$\hat{Y}_{0.05}/Y$	$\hat{Y}_{0.20}/Y$
-0.3	Average of estimates	1.004	1.003	1.001	1.002	1.002	1.001
	Standard deviation of estimates	0.077	0.073	0.062	0.042	0.040	0.035
	Average of estimated standard error	0.076	0.072	0.061	0.041	0.039	0.033
	Coverage	0.953	0.951	0.949	0.946	0.942	0.942
0	Average of estimates	1.013	1.012	1.008	1.001	1.001	1.001
	Standard deviation of estimates	0.080	0.070	0.059	0.041	0.039	0.033
	Average of estimated standard error	0.081	0.072	0.060	0.040	0.038	0.033
	Coverage	0.951	0.954	0.951	0.944	0.942	0.946
0.5	Average of estimates	1.053	1.018	1.009	1.004	1.003	1.002
	Standard deviation of estimates	0.586	0.104	0.072	0.057	0.045	0.037
	Average of estimated standard error	0.233	0.094	0.070	0.051	0.045	0.036
	Coverage	0.950	0.951	0.945	0.950	0.955	0.955
1.0	Average of estimates	1.064	1.030	1.013	1.013	1.009	1.006
	Standard deviation of estimates	0.515	0.162	0.090	0.094	0.066	0.047
	Average of estimated standard error	0.277	0.128	0.086	0.070	0.059	0.046
	Coverage	0.930	0.929	0.930	0.946	0.949	0.951

Table 5
Population Total Subsampling Estimates Scaled by Y for Model 2

		$N = 300$			$N = 1,000$		
ρ_{xy}		\hat{Y}_0/Y	$\hat{Y}_{0.05}/Y$	$\hat{Y}_{0.20}/Y$	\hat{Y}_0/Y	$\hat{Y}_{0.05}/Y$	$\hat{Y}_{0.20}/Y$
-0.3	Average of estimates	1.010	1.009	1.006	1.003	1.003	1.002
	Standard deviation of estimates	0.098	0.092	0.078	0.052	0.049	0.041
	Average of estimated standard error	0.094	0.089	0.074	0.051	0.048	0.040
	Coverage	0.935	0.926	0.931	0.952	0.955	0.955
0	Average of estimates	1.008	1.007	1.005	1.002	1.002	1.001
	Standard deviation of estimates	0.065	0.062	0.050	0.034	0.032	0.027
	Average of estimated standard error	0.064	0.061	0.051	0.034	0.032	0.028
	Coverage	0.953	0.950	0.952	0.947	0.951	0.955
0.5	Average of estimates	1.002	1.002	1.001	1.001	1.001	1.001
	Standard deviation of estimates	0.035	0.033	0.028	0.019	0.018	0.015
	Average of estimated standard error	0.035	0.034	0.029	0.019	0.018	0.016
	Coverage	0.954	0.950	0.951	0.965	0.967	0.965
1.0	Average of estimates	1.001	1.001	1.001	1.000	1.000	1.000
	Standard deviation of estimates	0.021	0.020	0.017	0.012	0.011	0.010
	Average of estimated standard error	0.021	0.020	0.017	0.012	0.011	0.009
	Coverage	0.952	0.949	0.954	0.947	0.947	0.943

For Model 2, the inclusion probability is a function of the covariate. As a result, \hat{N}_{CH} is not an appropriate estimator for N . We observe that \hat{N}_{CH} significantly underestimates the true population size. On the other hand, \hat{N}_{p_A} provides a good estimate of N . The bias in \hat{N}_{p_A} decreases as p_A increases in Model 2. Further, the relative bias decreases as the population size increases.

As expected, the standard deviation of \hat{N}_{p_A} decreases as the area frame inclusion probability p_A increases. For example, in Model 1 where $N = 300$, the inclusion of a 5% area frame sample reduces the relative standard deviation from 0.077 to 0.059, a 23% reduction. When a 20% area frame sample is utilized, the relative standard deviation decreases from 0.077 to 0.048, a 38% reduction. When $N = 1,000$, the inclusion of a 5% area frame sample decreases the relative standard deviation from 0.035 to 0.030, a 14% reduction. Increasing the area frame sample to 20% reduces the relative standard deviation from 0.035 to 0.025, a decrease of 29%. Generally speaking, the relative standard errors decrease as population size increases. Although the average of the estimated standard error of \hat{N}_{p_A} is smaller than the empirical standard deviation, the difference is relatively small. Also, the coverage probabilities of the 95% confidence interval based on \hat{N}_{p_A} are very close to 0.95. In contrast, the coverage probabilities of the 95% confidence interval based on \hat{N}_{CH} are 0.271 and 0.009 for $N = 300$ and $N = 1,000$, respectively.

Based on our simulations, we recommend the use of \hat{N}_{p_A} with a large value of p_A . The choice of p_A is determined in practice by area frame sampling costs, which are not taken into consideration in our study.

3.7.2 Population Total Estimation

For population totals, we observe results that are very similar to what we observed for the population size. In general, relative biases and standard errors decrease as p_A increases and as the population size increases. We also notice that the average relative estimated standard error is very close to the empirical standard deviation of the standardized estimator standardized by the total. This suggests that the approximate standard error formula in (7) is a good estimate of the standard error. Note also that the empirical coverage probabilities are mostly within three standard errors of 0.95. That is, most of the empirical coverage probabilities fall within $(0.95 \pm 3[(0.95 \times 0.05/1,000)^{1/2}]) = (0.929, 0.971)$.

4. SUMMARY

In this paper, we studied the performance of the estimated Horvitz-Thompson estimator of the population size and total based on samples from area and list frames. We presented methods for estimating the parameters of the logistic regression model for the inclusion probabilities. Though numerous models and other estimators are considered in Haines (1997), we presented simulation study results for only two models and a few estimators.

We believe the methods used in this paper are potentially very useful to survey researchers because list frame incompleteness is a fact of life. Our results are among the first to suggest a method of estimating population totals which account for incompleteness and model the inclusion probabilities as a function of the covariates.

ACKNOWLEDGEMENTS

The authors thank the editor and an associate editor for their comments which improved the content and presentation of this paper. We also wish to thank Christine Bunck, BEST Program Manager, Biological Resources Division, U.S. Geological Survey, for financial support of this research through a research work order to North Carolina State University. The views expressed are attributed to the authors and do not necessarily reflect those of the Census Bureau.

REFERENCES

- ALHO, J.M. (1990). Logistic regression in capture-recapture models. *Biometrics*, 46, 623-635.
- BOSECKER, R.R., and FORD, B.L. (1976). Multiple frame estimation with stratified overlap domain. *Proceedings of the Social Statistics Section, American Statistical Association*, 219-224.
- BURNHAM, K.P. (1972). Estimation of Population Size in Multiple Capture Studies when Capture Probabilities Vary Among Animals. Ph.D. thesis, Oregon State University.
- BURNHAM, K.P., and OVERTON, W.S. (1978). Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika*, 65, 625-633.
- BURNHAM, K.P., and OVERTON, W.S. (1979). Robust estimation of population size when capture probabilities vary among animals. *Ecology*, 60, 927-936.
- CHAO, A. (1988). Estimating animal abundance with capture frequency data. *Journal of Wildlife Management*, 52, 295-300.
- CHAO, A., LEE, S.-M., and JENG, S.-L. (1992). Estimating population size for capture-recapture data when capture probabilities vary by time and individual animal. *Biometrics*, 48, 201-216.
- CHAPMAN, D.G. (1951). Some Properties of the Hypergeometric Distribution with Applications to Zoological Censuses. University of California, University of California Publication in Statistics.
- COCHRAN, R.S. (1965). Theory and Applications of Multiple Frame Surveys. Ph.D. thesis, Iowa State University.
- COCHRAN, W.G. (1977). *Sampling Techniques*. 3rd edition. New York: John Wiley & Sons.
- COWAN, C.D., and MALEC, D. (1986). Capture-recapture models when both sources have clustered observations. *Journal of the American Statistical Association*, 81, 347-353.

- FAULKENBERRY, G.D., and GAROUI, A. (1991). Estimating a population total using an area frame. *Journal of the American Statistical Association*, 86, 445-449.
- FECOSO, R., TORTORA, R.D., and VOGEL, F.A. (1986). Sampling frames for agriculture in the United States. *Journal of Official Statistics*, 2, 279-292.
- FIENBERG, S.E. (1992). Bibliography on capture-recapture modelling with application to census undercount adjustment. *Survey Methodology*, 18, 143-154.
- FULLER, W.A., and BURMEISTER, L.F. (1972). Estimators for samples selected from two overlapping frames. *Proceedings of the Social Statistics Section, American Statistical Association*, 245-249.
- HAINES, D.E. (1997). Estimating Population Parameters Using Multiple Frame and Capture-Recapture Methodology. Ph.D. thesis, North Carolina State University.
- HAINES, D.E., and POLLOCK, K.H. (1998a). Combining multiple frames to estimate population size and totals. *Survey Methodology*, 24, 79-88.
- HAINES, D.E., and POLLOCK, K.H. (1998b). Estimating the number of active and successful bald eagle nests: an application of the dual frame method. *Environmental and Ecological Statistics*, 5, 245-256.
- HANSEN, M.H., HURWITZ, W.N., and MADOW, W.G. (1953). *Sample Survey Methods and Theory I*. New York: John Wiley & Sons.
- HARTLEY, H.O. (1962). Multiple frame surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 203-206.
- HARTLEY, H.O. (1974). Multiple frame methodology and selected applications. *Sankhyā*, 36, 3, C, 99-118.
- HORVITZ, D.G., and THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- HUGGINS, R.M. (1989). On the statistical analysis of capture experiments. *Biometrika*, 76, 133-140.
- KOTT, P.S., and VOGEL, F.A. (1995). Multiple-frame business surveys. *Business Survey Methods* (Ed. B.G. Cox.). New York: John Wiley & Sons. 185-203.
- LINCOLN, F.C. (1930). Calculating Waterfowl Abundance on the Basis of Banding Returns. U.S. Department of Agriculture, Circular, 118.
- LUND, R.E. (1968). Estimators in multiple frame surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 282-288.
- NEALON, J.P. (1984). Review of the Multiple and Area Frame Estimators, Staff Report 80. U.S. Department of Agriculture, Statistical Reporting Service, Washington, D. C.
- NORRIS, J.L., and POLLOCK, K.H. (1996). Nonparametric MLE under two closed capture-recapture models with heterogeneity. *Biometrics*, 52, 639-649.
- OTIS, D.L., BURNHAM, K.P., WHITE, G.C., and ANDERSON, D.R. (1978). Statistical inference for capture data on closed animal populations. *Wildlife Monographs*, 62, 1-135.
- PETERSEN, C.G.J. (1896). The yearly immigration of young plaice into the Limfjord from the German Sea, *Rep. Danish Biol. Sta.*, 6, 1-48.
- POLLOCK, K.H. (1991). Modeling capture, recapture, and removal statistics for estimation of demographic parameters for fish and wildlife populations: past, present, and future. *Journal of the American Statistical Association*, 86, 225-238.
- POLLOCK, K.H., HINES, J.E., and NICHOLS, J.D. (1984). The use of auxiliary variables in capture-recapture and removal experiments. *Biometrics*, 40, 329-340.
- POLLOCK, K.H., TURNER, S.C., and BROWN, C.A. (1994). Use of capture-recapture techniques to estimate population size and population totals when a complete frame is unavailable. *Survey Methodology*, 20, 117-124.
- QUENOUILLE, M.H. (1956). Notes on bias in estimation. *Biometrika*, 43, 353-360.
- SEKAR, C.C., and DEMING, W.E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44, 101-115.
- WOLTER, K.M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, 81, 338-346.
- WOLTER, K.M. (1990). Capture-recapture estimation in the presence of a known sex ratio. *Biometrics*, 46, 157-162.

An Estimation Method for Nonignorable Nonresponse

JEAN-FRANÇOIS BEAUMONT¹

ABSTRACT

When a survey response mechanism depends on a variable of interest measured within the same survey and observed for only part of the sample, the situation is one of nonignorable nonresponse. In such a situation, ignoring the nonresponse can generate significant bias in the estimation of a mean or of a total. To solve this problem, one option is the joint modelling of the response mechanism and the variable of interest, followed by estimation using the maximum likelihood method. The main criticism levelled at this method is that estimation using the maximum likelihood method is based on the hypothesis of error normality for the model involving the variable of interest, and this hypothesis is difficult to verify. In this paper, the author proposes an estimation method that is robust to the hypothesis of normality, so constructed that there is no need to specify the distribution of errors. The method is evaluated using Monte Carlo simulations. The author also proposes a simple method of verifying the validity of the hypothesis of error normality whenever nonresponse is not ignorable.

KEY WORDS: Nonignorable nonresponse; Maximum likelihood; Estimation equations; Regression imputation; Reweighting.

1. INTRODUCTION

When a survey response mechanism depends on a variable of interest measured in the same survey and observed for only part of the sample, the situation is one of nonignorable nonresponse. In measuring income, for example, it may be realistic to assume that low income earners will exhibit a lower tendency to respond than high income earners, or vice versa. Readers will find in Little (1982) a formal definition of the concept of nonignorable nonresponse. In such a situation, ignoring the nonresponse can generate significant bias in the estimation of a mean or of a total. To solve this problem, one option is the joint modelling of the response mechanism and the variable of interest, followed by estimation using the method of maximum likelihood, used for example in Greenlees, Reece and Zieschang (1982), and imputation of the missing values. The main criticism levelled at this method is that estimation using the method of maximum likelihood is based on the hypothesis of error normality for the model involving the variable of interest, and this hypothesis is difficult to verify.

Rancourt, Lee and Särndal (1994) described simple correction factors aimed at reducing the bias generated by nonresponse that is not ignorable without reference to the hypothesis of normality and in the absence of a response mechanism model. These correction factors, however, are only available for ratio imputation.

In this paper, the author proposes an estimation method that is robust with respect to the hypothesis of normality, so constructed that there is no need to specify the distribution of errors. The author also proposes a simple method of verifying the validity of the hypothesis of error normality whenever nonresponse is not ignorable.

In section 2, the problem is defined and some notation is introduced. In section 3, various estimators of the mean of a population are introduced for a variety of hypotheses concerning the response mechanism and the distribution of data. In section 4, an estimation method is proposed for nonignorable nonresponse. In section 5, the author describes the results of a simulation study used to compare the estimators described in the two preceding sections. Finally, the last section contains a brief discussion.

2. NOTATION

In the following, we attempt to estimate the mean of a variable Y for a certain population P . To do so, we select a sample S , and the variable Y is observed for only part of the sample. The sample of respondents is denoted R , and the sample of nonrespondents is denoted O . We assume that there is at least one variable that is observed for all the sampling units and correlated with Y .

The estimator of the mean, $\mu = \sum_{i \in P} Y_i / N$, where N is the size of the population, can be obtained by weighting the respondent units:

$$\mu_P^* = \frac{\sum_{i \in R} w_i w_{R,i}^* Y_i}{\sum_{i \in R} w_i w_{R,i}^*}, \quad (2.1)$$

where w_i denotes the sampling weights that correspond to the inverse selection probability and $w_{R,i}^*$ denotes the weights that correspond to the estimated inverse response probability. Another estimator of the mean can be obtained by imputing the missing values:

¹ Jean-François Beaumont, Household Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

$$\mu_i^* = \frac{\sum_{i \in R} w_i Y_i + \sum_{i \in O} w_i Y_i^*}{\sum_{i \in S} w_i}, \quad (2.2)$$

where Y_i^* denotes values that are imputed for the non-respondent units.

For the sake of simplicity, we assume, in the following, that the sampling weights are constant for all units of the population. Thus, we can eliminate w_i from equations (2.1) and (2.2). We also assume that there is only one observed variable for all sampling units. This variable is denoted X .

3. CURRENT ESTIMATION METHODS

In this section, equations (2.1) and (2.2) are developed under a variety of hypotheses concerning the response mechanism and the distribution of data, and appropriate estimation methods are described. In section (3.1), we assume a uniform response mechanism; in section (3.2), we assume a response mechanism that depends on X , while in section (3.3), we assume a response mechanism that depends on Y . The response mechanisms in sections (3.1) and (3.2) are ignorable, whereas the one in section (3.3) is not ignorable.

3.1 Uniform Response Mechanism

Assuming a uniform response mechanism, we have the same response probability for all sampling units. Thus, estimator (2.1) becomes:

$$\mu_{P,U}^* = \frac{\sum_{i \in R} Y_i}{n_R}, \quad (3.1)$$

where n_R is the total number of respondents. This estimator is the very same one we would have obtained by using equation (2.2) and by imputing the respondent mean for all nonrespondents.

3.2 Response Mechanism Dependent on X

When the response mechanism depends on variable X (correlated with Y), estimator (3.1) might be strongly biased. It is then preferable to use this variable as additional information for the estimation of mean μ .

Estimator (2.1) can be obtained by replacing $1/w_{R,i}$ by the estimated response probability using a logistic regression. A response probability model is therefore needed. If we only have one observed variable (X) for all sampling units, the model can be written as follows:

$$P(R_i = 1 | X_i) = \frac{1}{1 + \exp[-(\alpha_0 + \alpha_1 X_i)]},$$

where α_0 and α_1 are parameters to be estimated (using the maximum likelihood method, for example) and R_i is a

dichotomous variable equal to 1 if unit i responds and to 0 otherwise. The estimator of the mean obtained in this way is denoted $\mu_{P,X}^*$.

If we prefer to use estimator (2.2) instead, the missing values can be imputed using the following model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad (3.2)$$

where β_0 and β_1 are unknown parameters and ε_i is a random error term of zero mean that is not correlated with X_i . The imputed values are given by: $Y_i^* = B_0^* + B_1^* X_i$, where B_0^* and B_1^* are estimates (obtained by means of the method of least squares using the respondent units) of B_0 and B_1 which are in turn estimates of β_0 and β_1 . In fact, B_0 and B_1 are the estimates that would have been obtained (using the method of least squares) if we had observed all the units of sample S . The estimator obtained in this way is denoted $\mu_{I,X}^*$.

Note that all the models considered in this document are assumed to be valid for all the units of sample S .

We could also add a residual to the imputed values in order to better estimate the variance due to sampling (see for example Gagnon, Lee, Rancourt and Särndal 1996). However, this technique still does not make it possible to estimate the variance due to imputation. Moreover, it tends to produce estimates of the mean that are less precise than if no residual had been added. Since this paper does not deal with variance estimation, we have chosen not to add residuals to the imputed values. This has the added advantage of simplifying the calculation of $\mu_{I,X}^*$.

3.3 Response Mechanism Dependent on Y

All the estimators of the mean discussed so far can be strongly biased when the response mechanism depends on Y (nonignorable response mechanism). For such a response mechanism, the response probability can be modelled as follows:

$$P(R_i = 1 | Y_i) = \frac{1}{1 + \exp[-(\alpha_0 + \alpha_1 Y_i)]}. \quad (3.3)$$

Since variable Y is only observed for respondent units, it is impossible to obtain an estimate for α_0 and α_1 using the maximum likelihood method. Model (3.2) can also be used. However, the parameter estimates will not be consistent since $E(\varepsilon_i | R_i = 1)$ and $E(\varepsilon_i X_i | R_i = 1)$ are not zero. Even if we had consistent estimates, the missing values could not be imputed as described in section (3.2) since $E(Y_i | R_i = 0, X_i) \neq \beta_0 + \beta_1 X_i$ (Greenlees, Reece and Zieschang 1982). If, for example, the response probability correlates positively with the variable of interest Y , then, for a given value of X , the mean of nonrespondent units will be lower than that of respondent units, and will therefore be lower than the mean of all units taken together. A similar argument can be presented if the response probability correlates negatively with the variable of interest. In fact, it can be shown that

$$E(Y_i | R_i = 0, X_i) = \beta_0 + \beta_1 X_i - \frac{\text{cov}(Y_i, p(Y_i) | X_i)}{1 - E(p(Y_i) | X_i)},$$

where $p(Y_i) = P(R_i = 1 | Y_i)$.

The two approaches in section (3.2) are therefore invalid when the response mechanism is not ignorable. In such a situation, a better approach would be to estimate the parameters of models (3.2) and (3.3) simultaneously. The method of maximum likelihood can be used to this end. This method, however, requires as an additional hypothesis that errors ε_i follow a normal distribution (or any other distribution relevant to the type of data analyzed) with constant variance σ^2 , and that they be mutually independent. The natural logarithm of the likelihood function l can be written as follows:

$$l = \sum_{i \in R} \ln[p(Y_i)f(Y_i | X_i)] + \sum_{i \in O} \ln[1 - E(p(Y_i) | X_i)], \quad (3.4)$$

where $f(Y_i | X_i)$ is the probability density function of a normal distribution with a mean $\beta_0 + \beta_1 X_i$ and variance σ^2 . The method of maximum likelihood consists in finding the parameter values which maximize l . To carry out the maximization, it must be possible to approximate $E(p(Y_i) | X_i)$. This can be achieved by using a numerical integration method similar to that of Greenlees, Reece and Zieschang (1982). In this paper, the following approximation (Zeger, Liang and Albert 1988) has been used instead:

$$E(p(Y_i) | X_i) \approx \frac{1}{1 + \exp\{-k[\alpha_0 + \alpha_1(\beta_0 + \beta_1 X_i)]\}}, \quad (3.5)$$

where $k = 1/\sqrt{c^2 \sigma^2 \alpha_1^2 + 1}$ and $c = 16\sqrt{3}/15\pi$. This approximation is based on the hypothesis that errors follow a normal distribution with constant variance. This approximation was preferred to a method of numerical integration because it is simpler and computationally faster, an advantage that must be considered seriously before any simulation study is undertaken. Finally, equation (3.4) was maximized using the Newton-Raphson algorithm and the NLIN procedure of the SAS software (SAS Institute Inc. 1990).

Once the parameters of models (3.2) and (3.3) have been estimated, estimators of the mean (2.1) or (2.2) can be chosen. Estimator (2.1) is obtained by replacing $w_{R,i}$ by $1/p^*(Y_i)$, where $p^*(Y_i)$ is the estimated response probability. This estimator is denoted $\mu_{P,Y,ML}^*$. Estimator (2.2) can be obtained by determining imputed values Y_i^* in such a way that $\sum_{i \in S} e_i^2$ is minimized and that the constraints $\sum_{i \in R} e_i = 0$ and $\sum_{i \in S} e_i X_i = 0$ are met, where $e_i = Y_i - \beta_0 - \beta_1 X_i$, for $i \in R$, $e_i = Y_i^* - \beta_0^* - \beta_1^* X_i$, for $i \in O$, and β_0^* and β_1^* are the estimates of β_0 and β_1 respectively. The estimator of the mean can then be written as follows: $\mu_{I,Y,ML}^* = \beta_0^* + \beta_1^* \sum_{i \in S} X_i / n$, where n is the size of sample S .

The reasoning behind this approach is that the two previous constraints would have been met if variable Y had been observed for all units in the sample and if this variable had been modelled using model (3.2).

4. PROPOSED METHOD OF ESTIMATION

This section describes the proposed method of estimation for a nonignorable response mechanism (section 4.1), as well as a graphic method (section 4.2) that can be used to verify the error normality hypothesis of model (3.2).

4.1 Method of Estimation for a Response Mechanism Dependent on Y

The method of maximum likelihood is valid when errors exhibit a normal distribution and have the same variance. When this hypothesis does not hold, it is preferable to use a more robust method of estimation.

If response probabilities $p(Y_i)$ were known and greater than zero for all sampling units, a robust method of estimation (in terms of both the error normality hypothesis and model 3.2) would consist in minimizing the error sum of squares weighted by the inverse response probability $p(Y_i)$. This minimization is equivalent to solving the system of equations

$$\sum_{i \in R} \frac{1}{p(Y_i)} (Y_i - \beta_0 - \beta_1 X_i) Z_{ik} = 0, \quad k = 1, 2, \quad (4.1)$$

where $Z_{i1} = 1$ and $Z_{i2} = X_i$. This approach is considered robust with respect to the normality hypothesis since the method of least squares does not require that the distribution of errors be specified. Weighting by means of the inverse response probability also provides a certain robustness in terms of model (3.2). In fact, estimators B_0^* and B_1^* obtained using equation (4.1) are consistent with respect to the response mechanism for B_0 and B_1 (which are the estimators of β_0 and β_1 that would have been obtained if there had been no nonresponse) regardless of the validity of the model. A similar argument may be found in Särndal, Swensson and Wretman (1992, p. 519), but in terms of the sample selection mechanism instead of the response mechanism.

Likewise, if the probability density function $f(Y_i | X_i)$ was known (not necessarily normal and yet not dependent on the parameters of model 3.3), we could then estimate parameters α_0 and α_1 of model (3.3) using the maximum likelihood method, for example, and solve the system of equations

$$\sum_{i \in R} \frac{\partial}{\partial \alpha_k} \ln[p(Y_i)] + \sum_{i \in O} \frac{\partial}{\partial \alpha_k} \ln[1 - E(p(Y_i) | X_i)] = 0, \quad (4.2)$$

for $k = 0$ and $k = 1$.

Thus, the estimates of parameters β_0 , β_1 , α_0 and α_1 are obtained by solving the unbiased estimation equations (4.1) and (4.2). An algorithm that can be used to find the solution consists in solving alternately the systems of equations (4.1) and (4.2) until convergence is achieved. This requires the possibility of calculating $E(p(Y_i)|X_i)$ in equation (4.2). However, this last expectation requires that the distribution of errors ϵ_i be known, and in all likelihood it is unknown. To get around this problem, we must use an approximation, and a number of them can be considered, including approximation (3.5). Another option would be to develop a strategy based on the bootstrap method by selecting the respondent units proportionally to their inverse response probability. However, this method requires considerable computer processing time, and is not considered in this paper. Instead, we have chosen the following approximation, obtained by linearizing $p(Y_i)$ using a Taylor series assessed at point $E(Y_i|X_i)$ and by taking the expectation of the first two terms in this series:

$$E(p(Y_i)|X_i) \approx p(E(Y_i|X_i)) = p(\beta_0 + \beta_1 X_i). \quad (4.3)$$

It should be noted that the expectation of the second term in the series is zero. This approximation offers the advantage of requiring only the first moment of the distribution of Y_i conditional on X_i . In this sense, it should be robust with respect to the error normality hypothesis since it does not require that the error distribution be specified. Of course, if the distribution of errors is known or can be properly estimated, it will be possible to find better approximations than (4.3) although, in this case, it may be preferable to use the maximum likelihood method.

Another interesting property of approximation (4.3) is that alternately solving the systems of equations (4.1) and (4.2) might be achieved using the following algorithm:

1. determine initial values for the response probabilities (or for parameters α_0 and α_1), e.g., let $p(Y_i)^{(0)} = 1$ for all the respondent units;
2. let $j = 1$, where j is the number of iterations;
3. solve the system of equations (4.1) by means of the current response probability estimates, $p(Y_i)^{(j-1)}$, using a weighted regression procedure to obtain $\beta_0^{(j)}$ and $\beta_1^{(j)}$;
4. impute the missing values using $Y_i^{(j)} = \beta_0^{(j)} + \beta_1^{(j)} X_i$ for $i \in O$;
5. solve the system of equations (4.2) by using a logistic regression procedure to obtain $p(Y_i)^{(j)}$;
6. stop once convergence has been achieved, otherwise let $j = j + 1$ and return to step 3.

It is sufficient then to simply have a linear regression procedure and a logistic regression procedure to obtain the desired estimates. This algorithm is a very efficient means of finding the solution although, in certain cases, many iterations might be needed before convergence is achieved.

In actual practice, it did converge in all cases where it was used. It should also be noted that this algorithm shows certain similarities with the EM algorithm used by Dempster, Laird and Rubin (1977), except that here we do not maximize a likelihood function.

For the simulations in the next section, we selected instead the Newton-Raphson algorithm which converges more rapidly. However, the above-mentioned algorithm had to be used for the few cases in which the Newton-Raphson algorithm met with convergence problems.

The proposed algorithm might be very useful as a means of providing initial values for a more rapid algorithm such as the Newton-Raphson one. The proposed algorithm could simply be used with a not very demanding convergence criterion so that, after only a few iterations, it could provide sufficiently good initial values to ensure convergence of the Newton-Raphson algorithm. In a different context, Beaumont and Demnati (1998) used a similar approach by beginning the iterative process using an algorithm of the EM type so as to provide the initial values for a more rapid algorithm of the Newton-Raphson type. They were able to show empirically that the combination of the two algorithms represents a sound compromise between processing time and efficiency in finding a solution.

As in section (3.3), once the parameters of models (3.2) and (3.3) are estimated, we can select estimators of the mean (2.1) or (2.2). Estimator (2.1) is obtained by replacing $w_{R,i}^*$ by $1/p^*(Y_i)$, where $p^*(Y_i)$ is the estimated response probability. This estimator is denoted $\mu_{P,Y,ROB}$. Estimator (2.2) is also obtained as in section (3.3) by determining the imputed values Y_i^* in such a way that $\sum_{i \in S} e_i^2$ is minimized and the constraints $\sum_{i \in S} e_i = 0$ and $\sum_{i \in S} e_i X_i = 0$ are met, where $e_i = Y_i - B_0^* - B_1^* X_i$, for $i \in R$, and $e_i = Y_i^* - B_0^* - B_1^* X_i$, for $i \in O$. This estimator is denoted $\mu_{I,Y,ROB}$. The quality of these two estimators of the mean will depend largely on the validity of models (3.2) and (3.3) and on the quality of approximation (4.3).

A modification of step (5) for the algorithm presented in this section was proposed by Beaumont (1999). The results of a simulation study show that this modification provides results that are slightly better than those obtained using the method proposed in this paper. However, this no longer involves using the maximum likelihood method to estimate the parameters of model (3.3), given that $f(Y_i|X_i)$ is known and a logistic regression procedure can no longer be used for step (5). It should nevertheless be mentioned that it is not absolutely necessary to use the method of maximum likelihood to estimate α_0 and α_1 , although it is the method preferred in this paper.

4.2 Verifying the Error Normality Hypothesis

In order to use the method of maximum likelihood, we might be interested in verifying the error normality hypothesis (or rather the residual normality hypothesis since the errors are not observed). In the absence of nonresponse, a traditional method (D'Agostino 1986, p. 25, equation 2.11)

consists in producing the graph of $\Phi^{-1}[F_n^*(e_i)]$ in terms of residuals e_i , for $i \in S$, where $\Phi(\cdot)$ is the distribution function for a random variable having the standard normal distribution, and $F_n(\cdot)$ is the empirical distribution function. Whenever errors exhibit normal distribution, the points in this graph should more or less fall along a line having a slope $1/\sigma$ passing through the origin.

If there is nonresponse, the same strategy can be used as in the previous paragraph, but the empirical distribution function must be estimated using the respondent units. Since the units in the sample respond with unequal probabilities, the estimated empirical distribution function can be given by the formula (Särndal, Swensson and Wretman 1992, p. 199):

$$F_n^*(e_i) = \frac{\sum_{j: j \in R \text{ et } e_j \leq e_i} 1/p^*(Y_j)}{\sum_{j \in R} 1/p^*(Y_j)}.$$

Note that, in this last equation, the response probabilities are estimated as opposed to the Särndal, Swensson and Wretman formula, in which selection probabilities are known. Thus, the error normality hypothesis can be verified by producing the graph of $\Phi^{-1}[F_n^*(e_i)]$ in terms of residuals e_i , for $i \in R$. This method will be valid provided that $F_n^*(e_i)$ can correctly estimate $F_n(e_i)$, as is the case when the response probabilities are correctly estimated. When the nonresponse is not ignorable, and when the method of estimation proposed in this paper is used, the response probabilities should be properly estimated if models (3.2) and (3.3) are appropriate along with approximation (4.3).

5. SIMULATION STUDY

In order to compare the estimators of the mean presented in the two previous sections, we carried out a simulation study. We simulated 4 populations with a size of 1,000 according to model (3.2) with $\beta_0 = 2$ and $\beta_1 = 3$. Random variables X_i are independent of one another and they follow an exponential distribution of mean 1. Errors ϵ_i are independent of one another, are not correlated with the X_i and have a mean of zero and a variance σ^2 . In two populations, the errors follow a normal distribution ($\epsilon_i \sim \text{Nor}(0, \sigma^2)$), and in the other two populations, the errors follow an exponential distribution of mean σ recentred at $0(\epsilon_i \sim \text{Exp}(\sigma) - \sigma)$. For each of these distributions, one population has a standard deviation σ equal to 1.5 corresponding to a squared coefficient of correlation (between X and Y) of 80% ($R^2 = 80\%$), and the other has a standard deviation equal to 3 corresponding to a square coefficient of correlation of 50% ($R^2 = 50\%$).

For each population, we simulated 1,000 samples of respondents according to model (3.3) with $\alpha_1 = 0.5$. Parameter α_0 was determined separately for each of the 4 populations, so that the mean response rate would be 70%.

This parameter varied between -1.185 and -0.958. Note that we have here a census ($n = N = 1\,000$). The advantage of this is that we can concentrate solely on the nonresponse error since there is no sampling error. Moreover, the fact that populations of relatively large size (1,000) are generated makes it possible to emphasize the bias of the estimators instead of their variance, since the variance should diminish as the size of the population increases (for a fixed mean response rate).

For each of the 1,000 samples of respondents, we calculated the 7 estimates of the mean described in the two previous sections. We then calculated, for each population, the mean and the variance of these 1,000 estimates, denoted $\bar{\mu}^*$ and S_{μ}^{*2} , respectively. Finally, we calculated an estimate of the relative bias (expressed as a percentage), $RB^* = [(\bar{\mu}^* - \mu)/\mu] \times 100\%$, an estimate of the standard error associated with this relative bias, $SE^* = (100/\mu)\sqrt{S_{\mu}^{*2}/1\,000}$, and an estimate of the root mean square errors, $RMSE^* = \sqrt{S_{\mu}^{*2} + (\bar{\mu}^* - \mu)^2}$.

The results of the simulation study are shown in Table 1. An analysis of this table indicates that, regardless of the error distribution, the relative bias and the mean square error of all the estimators is lower when the correlation between X and Y is greater, which is not surprising.

Table 1
Simulation Results Used to Compare 7 Estimators of the Mean μ

Estimator	$R^2 = 80\%$			$R^2 = 50\%$		
	RB*(%)	SE*	RMSE*	RB*(%)	SE*	RMSE*
Population with normally distributed errors						
$\mu_{P,U}^*$	16.90	0.03	0.84	26.68	0.04	1.33
$\mu_{P,X}^*$	5.65	0.02	0.28	18.02	0.03	0.90
$\mu_{P,Y,ML}^*$	-0.14	0.03	0.05	1.27	0.10	0.17
$\mu_{P,Y,ROB}^*$	1.14	0.03	0.08	10.12	0.06	0.51
$\mu_{I,X}^*$	5.50	0.02	0.27	17.74	0.03	0.89
$\mu_{I,Y,ML}^*$	0.13	0.03	0.04	1.03	0.07	0.13
$\mu_{I,Y,ROB}^*$	0.64	0.03	0.05	7.53	0.06	0.39
Population with exponentially distributed errors						
$\mu_{P,U}^*$	17.83	0.04	0.86	26.60	0.05	1.29
$\mu_{P,X}^*$	5.44	0.02	0.26	16.06	0.04	0.78
$\mu_{P,Y,ML}^*$	-0.54	0.02	0.04	5.18	0.05	0.26
$\mu_{P,Y,ROB}^*$	1.31	0.02	0.07	7.43	0.03	0.36
$\mu_{I,X}^*$	5.19	0.02	0.25	15.41	0.03	0.75
$\mu_{I,Y,ML}^*$	-3.42	0.03	0.17	-25.47	0.05	1.23
$\mu_{I,Y,ROB}^*$	0.49	0.02	0.04	4.07	0.03	0.20

An analysis of the relative bias indicates that the method of maximum likelihood provides best results when the errors are normally distributed, followed by the robust estimation method described in section (4.1). Estimators which assume a nonignorable response mechanism have a lower relative bias than those which incorrectly assume an ignorable response mechanism. Among the latter estimators, the most biased is estimator $\mu_{P,U}^*$. For a given method, there is generally little difference between the

weighted estimator (2.1) and the estimator that includes imputed values (2.2). However, the latter must be given a slight advantage.

The conclusions in the previous paragraph always apply when errors are exponentially distributed, except that the robust estimation method becomes the best. This observation should not be surprising since the method of maximum likelihood is based on the error normality hypothesis. However, the weighted estimator $\mu_{P,Y,ML}^*$ remains slightly biased, and this is more difficult to explain.

The conclusions drawn from an analysis of the relative bias still apply when analyzing the mean square error. In fact, estimators which are very biased show a strong tendency to having a high mean square error and vice versa.

6. DISCUSSION

When the hypothesis of a nonignorable response mechanism is realistic, and when the hypothesis of error normality for linear regression model (3.2) is justified, using the method of maximum likelihood may be appropriate. However, when the latter hypothesis is not justified, the results of the simulation study described in section 5 show that the robust estimation method presented in this paper is preferable.

Moreover, Beaumont (1999) described the results of another simulation study indicating that the estimation method proposed in this paper is robust with respect to both the error normality hypothesis and model (3.2). As for the method of maximum likelihood, it has been shown to be even more sensitive to the validity of model (3.2) than to the hypothesis of error normality. The latter method should therefore only be used when all the hypotheses associated with models (3.2) and (3.3) are reasonable.

Obviously, all estimators show little bias when non-response is very low. Likewise, when the coefficient of correlation between X and Y is very high, all estimators show little bias, except for the estimator which assumes a uniform response mechanism $\mu_{P,U}^*$. In either case, the choice of an estimator should be based on the criterion of simplicity, which favours the estimators in section (3.2), specifically estimator $\mu_{I,X}^*$.

It should be noted that models (3.2) and (3.3) could be complexified according to the nature of the problem. For example, other independent variables could be included in these models. Variable Y could also be categorized using dummy variables, and these dummy variables could be used in model (3.3) instead of variable Y itself.

In this paper, we have dealt only with the problem of the estimation of a mean when the response mechanism is not ignorable. However, the methods described in sections 3 and 4 apply to other types of estimation. For example, weights or imputed values could be used for the estimation of parameters in a given regression.

This paper has attempted to describe a robust estimation method with respect to the hypothesis of error normality for model (3.2), making it possible to reduce the bias due to a nonignorable response mechanism. In some future work, it would be interesting to evaluate simple methods of variance estimation using imputed data and this robust estimation method.

ACKNOWLEDGEMENTS

The author wishes to thank the Small Area and Administrative Data Division of Statistics Canada, which made this work possible. He also wishes to thank Eric Rancourt, the two referees as well as the associate editor for some useful comments which helped improve the quality of this paper.

REFERENCES

- BEAUMONT, J.-F., and DEMNATI, A. (1998). Parameter estimation for a finite mixture of distributions for dichotomous longitudinal data: comparing algorithms. *Proceedings, Symposium 98, Longitudinal Analysis for Complex Survey, Statistics Canada*, 191-197.
- BEAUMONT, J.-F. (1999). A robust estimation method in the presence of nonignorable nonresponse. *Proceedings of the Section on Survey Research Methods, American Statistical Association*. (To appear).
- D'AGOSTINO, R.B. (1986). Graphical analysis. *Goodness-of-fit Techniques*, (R.B. D'Agostino and M.A. Stephens, Ed.), 7-62. New York: Marcel Dekker.
- DEMPSTER, A.P., LAIRD, N.M., and RUBIN, R.B. (1977). Maximum likelihood from incomplete data via the EM-algorithm. *Journal of the Royal Statistical Society B*, 39, 1-38.
- GAGNON, F., LEE, H., RANCOURT, E., and SÄRNDAL, C.-E. (1996). Estimating the variance of the generalized regression estimator in the presence of imputation for the Generalized Estimation System. *1996 Proceedings of the Survey Methods Section, Statistical Society of Canada*, 151-156.
- GREENLEES, J.S., REECE, W.S., and ZIESCHANG, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, 77, 251-261.
- LITTLE, R.J.A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.
- RANCOURT, E., LEE, H., and SÄRNDAL, C.-E. (1994). Bias corrections for survey estimates from data with ratio imputed values for confounded nonresponse. *Survey Methodology*, 20, 137-147.
- SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- SAS INSTITUTE INC. (1990). *SAS/STAT User's Guide*, 2, Version 6, Fourth Edition. Cary, NC: SAS Institute Inc.
- ZEGER, S.L., LIANG, K., and ALBERT, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 44, 1049-1060.

An Approximate Design Effect for Unequal Weighting When Measurements May Correlate With Selection Probabilities

BRUCE D. SPENCER¹

ABSTRACT

It is common practice to estimate the design effect due to weighting by 1 plus the relative variance of the weights in the sample. This formula has been justified when the selection probabilities are uncorrelated with the variable of interest. An approximation to the design effect is provided to accommodate the situation in which correlation is present.

KEY WORDS: Weighting; Deff; Sampling variance; Complex samples.

1. INTRODUCTION

It is common practice to weight observations in an unequal probability sample by the reciprocals of selection probabilities. The rationale is that failure to use the weights will cause bias if the sampling weights correlate with the variable of interest. A drawback to weighting is an increase in sampling variance when the weights vary excessively in the sample. This increase may be quantified by the design effect. The design effect is the ratio of the variance of the statistic of interest under the design of interest to the variance of the statistic under simple random sampling with the same sample size (Kish 1965). Design effects are important both for approximating standard errors after the sample is in hand and for predicting standard errors ahead of time, which is critical for efficient design of samples.

Kish (1965, 1992) discussed an approximation for the design effect for weighted estimates from unequal probability samples: $1 + rvw$, with rvw defined as the relative variance of the weights in the sample. Thus, if w_i is the weight of unit i in the sample and \bar{w} is the sample mean, $rvw = n^{-1} \sum_{i=1}^n (w_i - \bar{w})^2 / \bar{w}^2$. Gabler, Haeder, and Lahiri (1999) used a superpopulation model to derive a design effect when clustering is present as well. Their formula, which agrees with design-based results in Kish (1965), reduces to $1 + rvw$ when there is zero intraclass correlation. The $1 + rvw$ approximation for the design effect is based on a model or design in which the weights are uncorrelated with the variable of interest (and hence an unweighted estimate would serve as well or better than the weighted estimate). Here we develop an approximation to the design effect under a model in which correlation may be present. In developing the approximation we do not assume that the population is sampled from a superpopulation. The accuracy of the approximation depends only on the characteristics of the sample design and the population of interest.

For simplicity, we will discuss single-stage unequal probability sampling with replacement. Heuristic extension of the results to sampling without replacement is indicated in section 4.

2. REGRESSION REPRESENTATION OF POPULATION AND SAMPLE DESIGN

Let y_i denote the measurement of interest, P_i the (draw-by-draw) selection probability for a sample of size n , and $w_i = 1/(nP_i)$ the sampling weight for unit i in a population of size N , $1 \leq i \leq N$. Observe that $\bar{P} = \sum_{i=1}^N P_i / N = N^{-1}$. Consider the least-squares population regression line

$$y_i = \alpha + \beta P_i + \varepsilon_i, \quad (1)$$

with $\alpha = \bar{Y} - \beta \bar{P}$, $\beta = \sum_{i=1}^N (y_i - \bar{Y})(P_i - \bar{P}) / \sum_{i=1}^N (P_i - \bar{P})^2$, and $\bar{Y} = \sum_{i=1}^N y_i / N$. Denote the population variances of the y 's, the ε 's, the ε^2 's, and the w 's by σ_y^2 , σ_ε^2 , $\sigma_{\varepsilon^2}^2$, and σ_w^2 , with, for example, $\sigma_y^2 = \sum_{i=1}^N (y_i - \bar{Y})^2 / N$. Denote the population correlation between y and P by $\rho_{y,P}$, between ε and w by $\rho_{\varepsilon,w}$, and between ε^2 and w by $\rho_{\varepsilon^2,w}$. It follows from the properties of least-squares, or equivalently from the definitions of α and β , that $\sum_{i=1}^N \varepsilon_i P_i = \sum_{i=1}^N \varepsilon_i / N = 0$ and $\sigma_\varepsilon^2 = (1 - \rho_{y,P}^2) \sigma_y^2$. If data are available, we can fit the regression representation (1) and estimate α , β , σ , and ρ by, say, $\hat{\alpha}$, $\hat{\beta}$, $\hat{\sigma}$, and $\hat{\rho}$.

Let $\hat{Y} = \sum_{i=1}^n w_i y_i$ denote the usual weighted estimator of the population total, Y . The variance of \hat{Y} is well-known (Cochran 1977, 253) to be

$$V(\hat{Y}) = n^{-1} \sum_{i=1}^N P_i (y_i / P_i - Y)^2. \quad (2)$$

Using the regression formulation (1), we may re-express the variance as

$$V(\hat{Y}) = \alpha^2 N(\bar{W} - N/n) + (1 - \rho_{y,P}^2) \sigma_y^2 N \bar{W} + N \rho_{\varepsilon^2,w} \sigma_{\varepsilon^2} \sigma_w + 2\alpha N \rho_{\varepsilon,w} \sigma_\varepsilon \sigma_w, \quad (3)$$

where $\bar{W} = \sum_{i=1}^N w_i / N$.

This expression does not rest on any assumptions about the fit of the regression model. (See section 5 for derivation).

If the regression model fits well enough so that $\rho_{\varepsilon^2,w}$ and $\rho_{\varepsilon,w}$ are zero, then the variance in (3) simplifies to $V(\hat{Y}) = \alpha^2 N(\bar{W} - N/n) + (1 - \rho_{y,P}^2) \sigma_y^2 N \bar{W}$. If simple random sampling

¹ Bruce D. Spencer, Department of Statistics and Institute for Policy Research, 2006 Sheridan Road, Northwestern University, Evanston, IL 60208, U.S.A.

with replacement had been used, the variance would have been $n^{-1}N^2\sigma_y^2$. Therefore, if $\rho_{\epsilon^2, w}$ and $\rho_{\epsilon, w}$ are negligible, the design effect is approximately

$$\text{deff} = (1 - \rho_{y, P}^2)n\bar{W}/N + (\alpha/\sigma_y)^2(n\bar{W}/N - 1). \quad (4)$$

This approximation does not require that the residuals from the regression are negligible, and it can hold when σ_ϵ is large. A referee has pointed out that the condition that $\rho_{\epsilon^2, w}$ and $\rho_{\epsilon, w}$ are negligible may seem unnatural in a model that regresses y on P rather than on $w \propto 1/P$. Note, however, that if we had not only zero correlation between ϵ and P but also independence, then we would have zero correlation between functions of ϵ and functions of P , and so $\rho_{\epsilon^2, w}$ and $\rho_{\epsilon, w}$ would be zero as well.

3. ESTIMATION OF DESIGN EFFECT

To estimate the design effect after the sample is in hand, we may use $1 + rvw$ to estimate $n\bar{W}/N$. To understand the rationale for this, note first that

$$1 + rvw = \frac{n^{-1} \sum_{i=1}^n w_i^2}{\bar{w}^2}. \quad (5)$$

The expectation of the numerator is $N\bar{W}/n$. The expectation of \bar{w} is N/n , and so the denominator of (5) may be taken as an estimator of $(N/n)^2$. Dividing the expectation of the numerator by $(N/n)^2$, we obtain $n\bar{W}/N$. Thus the design effect may be estimated from the sample by

$$(1 - \hat{\rho}_{y, P}^2)(1 + rvw) + (\hat{\alpha}/\hat{\sigma}_y)^2(rvw). \quad (6)$$

As a special case, note that if we set $\hat{\rho}_{y, P} = 0$, the case of "haphazard weighting" (Kish 1992), then the estimate of the design effect simplifies to

$$1 + rvw + rvw(\hat{\alpha}^2/\hat{\sigma}_y^2). \quad (7)$$

This estimate is close to Kish's approximation when $\hat{\alpha}/\hat{\sigma}_y$ is near zero.

4. SAMPLING WITHOUT REPLACEMENT

To derive the exact design effect for sampling without replacement would be more complex, as it would require consideration of joint selection probabilities for pairs of units. A heuristic extension of the results is easy, however. Recall that the ratio of the variance of a sample mean under simple random sampling without replacement to the variance under with-replacement sampling is approximately $(1 - n/N)$.

The results we have derived for the design effect will apply to single-stage unequal probability samples of n units without replacement if the variance of the Horvitz-Thompson estimator of the total is approximately $(1 - n/N)$ times the variance in (2), with P_i taken as n^{-1} times the overall selection probability for unit i (Särndal, Swensson, and Wretman 1992, 154).

5. DERIVATION OF VARIANCE FORMULA (3)

From (2) we have $V(\hat{Y}) = n^{-1}(\sum_{i=1}^N y_i^2/P_i - Y^2)$. Next, note that (1) implies that

$$Y^2 = (N\alpha + \beta)^2 = N^2\alpha^2 + 2N\alpha\beta + \beta^2 \quad (8)$$

and

$$\begin{aligned} \sum_{i=1}^N y_i^2/P_i &= \sum_{i=1}^N [\alpha^2/P_i + \beta^2 P_i + \epsilon_i^2/P_i + 2\alpha\beta + 2\alpha\epsilon_i/P_i + 2\beta\epsilon_i] \\ &= \alpha^2 \sum_{i=1}^N P_i^{-1} + \beta^2 + \sum_{i=1}^N \epsilon_i^2/P_i + 2N\alpha\beta + 2\alpha \sum_{i=1}^N \epsilon_i/P_i \\ &= \alpha^2 n \sum_{i=1}^N w_i + \beta^2 + n \sum_{i=1}^N \epsilon_i^2 w_i + 2N\alpha\beta + 2\alpha n \sum_{i=1}^N \epsilon_i w_i. \end{aligned} \quad (9)$$

Subtracting (8) from (9) and dividing by n yields

$$V(\hat{Y}) = \alpha^2 \left(\sum_{i=1}^N w_i - N^2/n \right) + \sum_{i=1}^N \epsilon_i^2 w_i + 2\alpha \sum_{i=1}^N \epsilon_i w_i.$$

To obtain (3), note that

$$\begin{aligned} \sum_{i=1}^N \epsilon_i^2 w_i &= N\rho_{\epsilon^2, w}\sigma_{\epsilon^2}\sigma_w + N\bar{W}\sigma_{\epsilon^2}^2 \\ &= N\rho_{\epsilon^2, w}\sigma_{\epsilon^2}\sigma_w + (1 - \rho_{y, P}^2)\sigma_y^2 N\bar{W} \end{aligned}$$

and

$$\sum_{i=1}^N \epsilon_i w_i = N\rho_{\epsilon, w}\sigma_{\epsilon}\sigma_w.$$

REFERENCES

- COCHRAN, W. G. (1977). *Sampling Techniques*. 3rd ed. New York: Wiley.
- GABLER, S., HAEDER, S., and LAHIRI, P. (1999). A model based justification of Kish's formula for design effects for weighting and clustering. *Survey Methodology*, 25, 1, 105-106.
- KISH, L. (1965). *Survey Sampling*. New York: Wiley.
- KISH, L. (1992). Weighting for unequal P_i . *Journal of Official Statistics*, 8, 2, 183-200.
- SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

On the Validity of Markov Latent Class Analysis for Estimating Classification Error in Labor Force Data

PAUL P. BIEMER and JOHN M. BUSHERY¹

ABSTRACT

The primary goal of this research is to investigate the validity of Markov latent class analysis (MLCA) estimates of labor force classification error and to evaluate the efficacy of MLC analysis as an alternative to traditional methods for evaluating data quality. We analyze interview data from the Current Population Survey (CPS) for the first three months of each of three years – 1993, 1995, and 1996 – and conduct an additional analysis of the CPS unreconciled reinterview data for approximately the same time periods. The reinterview data provides another approach for estimating CPS classification error that, when compared with the MLC estimates, helps to address the validity of the MLCA approach. Five dimensions of MLCA validity are addressed: (a) model diagnostics, (b) model goodness of fit across three years of CPS, (c) agreement between the model and test-retest reinterview estimates of response probabilities, (d) agreement between the model and test-retest reinterview estimates of inconsistency, and (e) the plausibility of patterns of classification error. In addition, we consider the robustness of the MLCA estimates to violations in the Markov assumption. Our analyses provides no evidence to question the validity of the MLC approach. The method performed well in all five validity tests.

KEY WORDS: Panel surveys; Nonsampling error; Unemployment; Data quality.

1. INTRODUCTION

The Current Population Survey (CPS) is a household sample survey conducted monthly by the U.S. Bureau of the Census to provide estimates of employment, unemployment, and other characteristics of the general U.S. labor force population. National estimates from the CPS of the size, composition, and changes in the composition of the labor force are published each month by the U.S. Bureau of Labor Statistics in Employment and Earnings. The CPS labor force estimates comprise one of the Nation's key economic indicators; since 1942, the Federal government has used the CPS data series to monitor month-to-month and year-to-year changes in labor force participation.

Given the importance of the CPS data series to public policy, there have been numerous evaluations of the accuracy of the data. For example, since the early 1950s, the Census Bureau has conducted the CPS Reinterview Program to evaluate the quality of the labor force data. The program involves drawing a small subsample (less than 5 percent) of the CPS respondents and re-asking some of the questions asked in the original interview – particularly the labor force questions. Until 1994, about one fourth of the sample received an unreconciled reinterview and three fourths received a reconciled reinterview. The reconciled reinterview component, which was used primarily for interview quality control purposes, was discontinued in 1994 due to concerns about the quality of the data. However, the unreconciled reinterview continues today and is used to estimate the test-retest reliability (or response consistency). Forsman and Schreiner (1991) provide a detailed description of the CPS Reinterview Program.

Several papers prepared by researchers outside the Census Bureau analyze the CPS Reinterview Program data to estimate the classification error in the CPS (*cf.* Sinclair and Gastwirth 1996, 1998; Biemer and Forsman 1992; Chua and Fuller 1987; Poterba and Summers 1986; Abowd and Zellner 1985). Recently, Poterba and Summers (1995) used data from the CPS Reinterview Program to estimate the CPS classification error rates and to evaluate the impact of classification error on labor market transition rates. As in the 1986 paper, their more recent analysis is based on the assumption that the CPS reinterview reconciliation process yields data which may be considered as the truth. Abowd and Zellner (1985) took similar approach.

Several authors (*viz.*, Sinclair and Gastwirth 1996, 1998; Biemer and Forsman 1992; Forsman and Schreiner 1991; Schreiner 1980) question the assumption that reconciled reinterview yields true values. They provide considerable evidence that the reinterview data are subject to substantial classification errors. In fact, this realization was responsible for the Census Bureau's decision to eliminate the reconciled reinterview portion of the CPS Reinterview Program in 1994.

As an alternative to the infallibility assumption, Chua and Fuller (1987) and Fuller and Chua (1985) apply a type of latent structure model to the CPS reconciled reinterview data to estimate the CPS response probabilities. For model identifiability, they impose tight restrictions on the response probabilities, forcing the bias due to classification error to be zero for both interview and reinterview. In addition, they assume independent classification errors for the interview and reinterview (referred to as the ICE assumption in the literature) and across the months in sample. The ICE

¹ Paul P. Biemer, Research Triangle Institute, Research Triangle Park, NC 27709; John M. Bushery, Bureau of Transportation Statistics, Washington, DC 20590, U.S.A.

assumption is a limitation of their analysis because evidence in the literature suggests that the assumption may not hold for the CPS (see, for example, O'Muircheartaigh 1991, and Singh and Rao 1995). Consequently, response probabilities estimated using the Chua and Fuller approach may be biased.

Sinclair and Gastwirth (1996) and Sinclair and Gastwirth (1998) apply a latent class modeling approach to the CPS interview-reinterview data using model restrictions originally proposed by Hui and Walter (1980) for medical diagnostic testing. Using the interview-reinterview data cross-classified by sex, Sinclair and Gastwirth assume that classification error probabilities are equal for males and females while labor force participation rates differ for these groups. Since the model parameters consume all the available degrees of freedom for parameter estimation, no residual degrees of freedom are available to test model lack-of-fit. Consequently, their analysis does not directly address whether these model assumptions hold for the CPS data.

In an examination of the determinants of rotation group bias, Shockey (1988) also applies latent class analysis to the CPS. His analysis suggests that the rotation group bias problem first reported by Bailer (1975) may be caused by response error arising from the interview administration. Shockey did not use reinterview data but rather relied on confirmatory factor analytic methods to support his claims. The sizes of his error rates were much larger than those reported by other authors which may be an indication of model bias. Unfortunately, like Sinclair and Gastwirth, Shockey's data set is not adequate to test fully the assumptions of the model he used.

The method of Markov latent class analysis, a promising approach for estimating the classification error in panel survey data, previously has not been applied to the CPS. This method takes advantage of the repeating nature of panel surveys to extract information on classification error directly from the interview data. The MLCA model is really a combination of two models: a latent Markov chain model representing the month to month transitions among the true labor force classifications and a classification error model representing the deviations from the true and observed labor classifications.

Because MLCA takes advantage of the repeating nature of panel surveys to extract information on classification error directly from the interview data, it does not require external, infallible measurements or remeasurements obtained by reinterview methods. In that regard, the method offers some advantages over both the Census Bureau's traditional methods and the methods of Chua and Fuller, Abowd and Zellner, Porterba and Summers, and Sinclair and Gastwirth for evaluating survey data quality in surveys. In many panel surveys, reinterviews are not feasible due to budget constraints, field work complexity, and respondent burden. MLCA may be the only way to assess the measurement error in these surveys. For panel surveys, such as the CPS, where reinterview data are

available, the reinterview and MLCA methods offer alternative analytical approaches for evaluating classification error. For example, as in the present analysis, MLCA can be used to model and test the traditional reinterview analysis assumptions. Further, MLCA analysis provides a statistical framework for combining the panel data and reinterview data to obtain even more information about classification error (van de Pol and Langeheine 1997).

Another advantage of MLCA is the potential for incorporating the entire panel data set into the estimates of classification error rather than only the relatively small sample selected for reinterview. As a result, a number of data quality issues for panel surveys that previously could not be explored for lack of data may now be tractable.

This paper reports our findings regarding the utility of the MLCA modeling approach for evaluating labor force classification error in the CPS. Software for fitting a wide variety of MLCA and other latent class models is available from several sources. The software employed in our analysis is *ℓEM* (Vermunt 1997), which can fit a large class of log-linear models with or without latent variables. The flexibility and generality of this software allow the measurement error analyst to test a considerable range of classification error models and to explore hypotheses regarding the causes and correlates of classification error.

In the next section, we describe the MLCA model and estimation methodology and its theoretical underpinnings. In section 3, we develop the MLCA methodology for the CPS application, fit a series of models to the CPS, and examine the fit of these models. In this section, we also produce estimates of classification error based upon the best MLCA model. In section 4, we conduct a number of tests of the validity of the MLCA estimates including a comparison of the MLCA estimates with those from new interview-reinterview analysis. Finally, in section 5, we summarize our findings and make recommendations regarding the utility of the MLCA method for future evaluations of labor force classification error.

2. MARKOV LATENT CLASS ANALYSIS FOR THREE TIME PERIODS

Markov latent class models were first proposed by Wiggins (1973) and refined by Poulsen (1982). Van de Pol and de Leeuw (1986) established conditions under which the model is identifiable and gave other conditions of estimability of the model parameters. In this section, we develop the MLCA model in the context of the CPS and suggest other applications and its generalizations.

Let the CPS target population be divided into L groups (such as age, race, or sex groups) and let the variable G be the label for group membership. For example, $G_i = 1$ if the i -th population member is in group 1, $G_i = 2$ for group 2 and so on. Let X_{gi} , Y_{gi} , and Z_{gk} denote the true labor force classifications for the i -th person in group

$G = g$ (for $g = 1, \dots, L$ and $i = 1, \dots, n_g$) where X_{gi} is defined as

$$X_{gi} = \begin{cases} 1 & \text{if person } (g, i) \text{ is employed} \\ & \text{in time period 1} \\ 2 & \text{if person } (g, i) \text{ is unemployed} \\ & \text{in time period 1} \\ 3 & \text{if person } (g, i) \text{ is not in the labor force} \\ & \text{in time period 1} \end{cases}$$

with analogous definitions for Y_{gi} and Z_{gi} for periods 2 and 3 respectively. Let $\pi_{x,y,z|g}$ denote $\Pr(X = x, Y = y, Z = z | G = g)$, let $\pi_{y|g,x}$ denote $\Pr(Y = y | X = x, G = g)$ and let $\pi_{z|g,y,x}$ denote $\Pr(Z = z | Y = y, X = x, G = g)$. Then, the probability that an individual in group g has labor status x in period 1, y in period 2, and z in period 3 is

$$\pi_{x,y,z|g} = \pi_{x|g} \pi_{y|g,x} \pi_{z|g,x,y} \quad (1)$$

Finally, under the first order Markov assumption, a necessary condition for model identifiability (see Van de Pol and de Leeuw 1986), we assume

$$\pi_{z|g,x,y} = \pi_{z|g,y} \quad (2)$$

i.e., at period 3, the true status of an individual does not depend on the period 1 status, once the period 2 status is known. An alternate interpretation is that the current status, given the prior period's status, does not depend upon the prior period's transition.

One can conceive of a number of scenarios where the Markov assumption may not hold for monthly labor force status. The assumption would be violated, for example, if individuals who are unemployed in period 2 are more likely to be unemployed in period 3, given they were also unemployed in period 1. The group of people unemployed in period 2 and period 1 probably includes a higher proportion of chronically unemployed people than the group unemployed in period 2 but not in period 1. That group (unemployed period 2, not period 1) likely contains a higher proportion of people temporarily out of work while changing jobs.

However, the validity of this assumption cannot be adequately explored using the observed data because the data are distorted to some unknown extent by the presence of classification errors. At least two methods for assessing the validity of the Markov assumption for panel data are available. Van de Pol and de Leeuw (1986) suggest a method based upon four waves of panel data that substitutes a second order Markov restriction for the first order restriction in (2). Another method, suggested by van de Pol and Langeheine (1997), uses a combination of labor force panel data and the reinterview data at each time period. Neither of these methods was employed in this paper to test the MLCA assumption directly. Instead, we assessed the overall validity of the MLCA estimates using the methods discussed in section 3.2 below. In section 3.6 we provide

some results from a simulation study to illustrate the robustness of the MLCA estimates of classification error to violations of the Markov assumption.

Now, consider the observed labor force classifications from the CPS denoted by A_{gi} , B_{gi} , and C_{gi} for periods 1, 2, and 3, respectively, where

$$A_{gi} = \begin{cases} 1 & \text{if person } (g, i) \text{ is classified as employed} \\ & \text{in time period 1} \\ 2 & \text{if person } (g, i) \text{ is classified as unemployed} \\ & \text{in time period 1} \\ 3 & \text{if person } (g, i) \text{ is classified as NLF} \\ & \text{in time period 1} \end{cases}$$

with analogous definitions for the response indicators, B_{gi} , and C_{gi} for periods 2 and 3, respectively. Using an extension of the notation established above, we denote the response probabilities in each of these classifications as $\pi_{a|g,x} = \Pr(A = a | G = g, X = x)$, with analogous definitions for $\pi_{b|g,y}$ and $\pi_{c|g,z}$. Thus, $\pi_{a=1|g,x=2}$ is the probability that the CPS classifies a person in group g as employed ($A = 1$) when the true status is unemployed ($X = 2$). Likewise, $\pi_{a=2|g,x=2}$ is the probability that the CPS correctly classifies a person in group g as unemployed.

Finally, we assume

$$\pi_{a,b,c|g,x,y,z} = \pi_{a|g,x} \pi_{b|g,y} \pi_{c|g,z} \quad (3)$$

or that classification error in the observed labor forces status is independent across the three months. This assumption, referred to as the local independence assumption, has been investigated for the CPS by Meyers (1988) in his review of the Abowd and Zellner (1985) estimation approach. Meyers concluded that the assumption "seems a reasonable approximation." Singh and Rao (1995), who studied the robustness of the assumption under a number of labor force population scenarios, reached a similar conclusion. Van de Pol and Langeheine (1997) modeled the joint distribution of panel data and reinterview data using latent class models to test for local independence for various types of labor force transitions. They found some evidence that people who change labor force status have lower reliability than those who do not, however the effect was quite small. Therefore, we shall also assume (3) without attempting any further investigation of its validity in this paper.

The CPS labor force classifications for each month of the first quarter of the year are the outcome variables in our analysis. Let A , B , and C denote the observed classifications and let X , Y , and Z denote the (unobserved) true classifications for January, February, and March, respectively. Let G denote some grouping (or stratification) variable to be defined later in the analysis. Under these assumptions, we can write the probability for classifying a CPS sample member in cell (g, a, b, c) of the $GABC$ table as follows:

$$\pi_{g,a,b,c} = \sum_{x,y,z} \pi_g \pi_{x|g} \pi_{a|g,x} \pi_{y|x,g} \pi_{b|y,g} \pi_{z|g,y} \pi_{c|g,z} \quad (4)$$

Extensions to more than one grouping variable are straightforward.

Under multinomial sampling, the likelihood function for the *GABC* table is

$$\Pr(GABC) = k \prod_{g,a,b,c} \pi_{g,a,b,c}^{n_{gabc}} \quad (5)$$

where k is the multinomial constant and Π denotes the product of the terms over the subscripts g , a , b , and c . Under the assumptions made previously, the model parameters are estimable using maximum likelihood estimation methods. Van de Pol and de Leeuw (1986) provides the formula for applying the E-M algorithm to estimate the parameters of this model and describes the conditions for their estimability. The *ℓ*EM software, applied to the CPS data sets in the next section, implements these methods.

3. APPLICATION TO THE CPS

3.1 Notation

Part of our evaluation of the MLCA approach will compare the MLCA estimates of classification error with estimates derived from the analysis of interview-reinterview data. Using the notation in the previous section, let A and A' denote the labor force classification for the original and reinterview, respectively, and define $\pi_a = \Pr(A = a)$ and $\pi_{a'} = \Pr(A' = a')$. Let AA' denote the observed interview-reinterview $K \times K$ cross-classification table and let $\pi^{AA'|X}$ denote the $K \times K$ matrix of cell probabilities, $\Pr(A = a, A' = a' | X = x)$. If we assume that $\pi_{aa'} = \Pr(A = a, A' = a' | X = x) = \pi_{a|x}^2$, referred to in the literature as the assumption of parallel measures (Bohrnstedt 1983), then

$$\pi^{AA'|X} = \pi^{A|X} (\pi^{A|X})^T \quad (6)$$

where $(\pi^{A|X})^T$ denotes the transpose of vector of conditional probabilities, $\pi^{A|X}$.

Let π^X denote the K -vector of true classification probabilities. Then

$$\pi^{AA'} = \pi^{AA'|X} \pi^X \quad (7)$$

i.e., the probability of the observed interview-reinterview classification table, $\pi^{AA'}$, is equal to the product of the matrix of conditional response probabilities, $\pi^{AA'|X}$, and the vector of true classification probabilities, π^X .

As described in the previous section, the MLCA of the CPS longitudinal data will provide maximum likelihood estimates of $\pi^{A|X}$ and π^X , allowing the estimation of $\pi^{AA'}$ via (6) and (7). We can estimate the test-retest reliability, R , for any labor force category by applying the usual estimation methods (see, for example, Bohrnstedt 1983) to this estimate of $\pi^{AA'}$. For our analysis, we compute the index

of inconsistency, $I = 1 - R$, which is the traditional reliability measure for CPS labor force data (see U.S. Bureau of the Census 1985). Let I_a denote the index of inconsistency for category $A = a$. Then an estimator of I_a is

$$\frac{gdr}{2\hat{\pi}_a(1 - \hat{\pi}_a)} \quad (8)$$

where gdr is the gross difference rate defined by

$$gdr_a = 2 \sum_{a \neq a'} \hat{\pi}_{a,a'} \quad (9)$$

and where $\hat{\pi}_a$ and $\hat{\pi}_{a,a'}$ denote latent class estimates of π_a and $\pi_{a,a'}$, respectively.

U.S. Bureau of the Census (1985, 88-91) provides the formulas for standard errors as well as an aggregate measure of inconsistency for all K categories combined, referred to as the aggregate index of inconsistency, I_{AG} . The aggregate index is a question-level measure of unreliability equal to $1 - \kappa$ (Hess, Singer and Bushery 2000) where κ is Cohen's kappa reliability measure (Cohen 1960) and is a weighted average of the category-level indexes.

Finally, given an estimate of π^X we can estimate the K -vector of measurement biases, denoted by β_A , associated with the K categories of A using the identity

$$\beta_A = \pi^A - \pi^X. \quad (10)$$

3.2 Assessing the Validity of the MLCA Methodology

The primary objective of this paper is to assess the validity of the MLCA approach. Previous research in the measurement of CPS classification error has not fully addressed the validity of the estimation approaches used (Meyers 1988). We hope to determine whether the MLCA approach is informative and useful for studying classification error in the CPS. In particular, we aim to determine whether the model estimates of error probabilities, $\pi^{A|X}$, reflect the actual levels of error in the CPS labor force classifications. Unfortunately, for the reasons mentioned previously, no generally accepted gold standard exists for assessing the accuracy of the CPS (see, for example, Sinclair and Gastwirth 1996, 1998, Biemer and Forsman 1992, and Schreiner 1980). Consequently, estimating the bias of MLCA estimates is not possible.

In what follows, we will investigate the validity of the MLCA estimates of CPS classification error using five criteria:

1. **Model diagnostics.** A necessary condition for model validity is that the model is plausible (*i.e.*, the assumptions are reasonable and are consistent with reality) and fits the data adequately. We use the traditional chi-square goodness of fit criteria and other diagnostic measures of model fit to assess the adequacy of the model specification and the degree to which the data are consistent with the model.

2. **Model Goodness of Fit Across Years of CPS.** An often-used technique for model validation is to assess the fit of the model for data that are independent of the data used for model building (see, for example, Kleinbaum, Kupper and Muller 1988, 330). This method is useful for avoiding model over-parameterization and data-driven (rather than theory-driven) model selection. In the present study, fitting the same model to data for each year separately is a form of this independent model verification technique. Model agreement across the years would tend to support the validity of the model structure. This method has a difficulty in the present application. After 1993, the CPS paper and pencil questionnaire was redesigned for Computer Assisted Personal Interview (CAPI) administration, so the magnitudes of the response errors may have changed after 1993. However, if the primary sources of response error in the CPS have not changed with the redesign, a model structure that adequately describes the error for 1993 should also describe the error for 1995 and 1996.
3. **Agreement of the MLCA Estimates and the Hui-Water Test-Retest Estimates of Response Probabilities.** The Hui-Walter (H-W) method (Hui and Walter 1980) for estimating CPS response probabilities uses unreconciled reinterview data (Sinclair and Gastwirth 1996; 1998). Although the MLCA and H-W methods both use latent class models, the model assumptions are very different. For example, the H-W method does not require the Markov assumption for model identifiability. Further, in this research, the data inputs to the H-W method are independent of those used for the MLCA method. Close agreement between the two sets of estimates supports the validity of both methods, while poor agreement suggests that at least one of the approaches is not valid. Strong agreement between the MLCA and H-W estimates also lend some assurance that the MLCA estimates of response probabilities are relatively robust to possible violations of the Markov assumption.
4. **Agreement of Model and Test-Retest Estimates of the Index of Inconsistency.** This criterion is similar to Criterion 3 because it compares estimates derived from MLCA with estimates based upon unreconciled reinterview data. However, this analysis does not rely on the validity of the Hui-Walter estimation methodology to assess MLCA estimation validity. Instead we use the MLCA estimates of classification error to compute estimates of the index of inconsistency using (7) to (9). We compare these estimates of reliability directly to the estimates of reliability from the CPS Reinterview Program, obtained from unreconciled reinterview data. Good agreement between the Reinterview and MLCA estimates supports the validity of both methods, while poor agreement

suggests that at least one of the approaches is not valid.

5. **Plausibility of Patterns of Classification Error.** Finally, the plausibility (or face validity) of the response probability estimates can also provide a test of validity. For example, it seems implausible that proxy responses to labor force questions should be more accurate than self-responses. Other patterns of classification error can also be reviewed and evaluated for plausibility. To the extent that the model estimates seem plausible, the face validity of the estimates is supported.

In the next section, we discuss our MLCA modeling results in the context of these criteria for validity. We begin with a description of the CPS data sets and the results of the model selection process.

3.3 The CPS Data Sets

In 1994, in conjunction with the implementation of computer assisted personal interviewing (CAPI), the CPS underwent a major redesign and a restructuring of the questions used to determine labor force status. Rothgeb (1994) provides a description of the CPS redesign. As a result of these improvements, we expect to see a difference (specifically a reduction) in classification error for the post-1994 CPS relative to 1993. Although not a primary objective of this research, we compared the error in the CPS before and after the redesign. We tested the MLCA approach for three years of the CPS – 1993, 1995, and 1996 – because the CPS unreconciled reinterview data were readily available for these time periods.

The CPS households are interviewed for four consecutive months, drop out of the survey for eight months, and then re-enter to be interviewed for a second series of four consecutive months. MLCA requires at least three consecutive interviews for identifiability of the model parameters. We had a choice of data sets which included all persons interviewed in three or four consecutive months of the CPS. Since using four months of data would reduce the sample size for the analysis by half, we chose to focus the analysis on three consecutive months – January, February, and March – for all three years of data. Nonresponse cases and cases where the whole household changed in one or more of the three months were excluded from the analysis.

The simplest MLCA model specifies that the response probabilities, $\pi_{a|x}$, $\pi_{b|y}$, and $\pi_{c|z}$, and the transition probabilities, $\pi_{y|x}$, $\pi_{z|y}$ are the same for all persons in the target population (referred to as homogeneity). However, our preliminary analysis (Biemer, Bushery and Flanagan 1997) indicated that response and transition probabilities were not homogeneous. To account for this heterogeneity, we explored a number of covariates and stratification variables for inclusion in the models, including: gender, education, mode of interview, proxy/self-response, and race. Of the

those considered, a variable derived from the CPS proxy/self response indicator best accounted for population heterogeneity. This variable, denoted by P , is defined as follows:

$$P = \begin{cases} 1 & \text{if all three interviews are conducted} \\ & \text{by self-response (SELF)} \\ 2 & \text{if two of the three interviews are conducted} \\ & \text{by self-response (MOSTLY SELF)} \\ 3 & \text{if two of the three interviews are conducted} \\ & \text{by proxy response (MOSTLY PROXY)} \\ 4 & \text{if all three interviews are conducted} \\ & \text{by proxy response (PROXY)} \end{cases}$$

Note, we now use P to represent the grouping variable, in place of G , which we used in section 2. Based upon previous research (for example, O'Muirheartaigh 1991), we expect that the Self group ($P=1$) to have less classification error than the Proxy group ($P=4$). We test this hypothesis as part of the estimate plausibility criterion (criterion 4 above).

The sample sizes for the three data sets used in our analysis are

1993:	45,291 persons
1995:	49,347 persons
1996:	41,751 persons

For 1993, approximately one-third of the sample is in the Self group, approximately one-fourth in the Proxy group, and the remaining sample members are distributed approximately equally between the Mostly Self and Mostly Proxy groups. For 1995 and 1996, slightly more sample members (one-third rather than one-fourth) are in the Proxy group.

3.4 Fitting the MLCA Models

To fit an MLCA model with a single grouping variable, P , the input data set was a $4 \times 3 \times 3 \times 3$ table of cell counts defined by the cross-classification of $P \times A \times B \times C$, where A , B , and C are the labor force classifications for January, February, and March, respectively.

The ℓ EM software and other software packages for fitting MLCA models assume simple random sampling, so the complex survey design of the CPS cannot be modeled exactly. It is possible to account for the unequal probability sampling structure of the CPS through the use of weighted and rescaled cell counts rather than the raw cell totals (Clogg and Eliason 1985). However, using unweighted data for the MLCA analysis affords two important advantages. First, we can compare the MLCA estimates with estimates from the previously cited studies on CPS classification error, all of which used unweighted data. Second, the CPS reinterview data used to assess Criteria 3 and 4 are unweighted and weights are not available. Consequently, at least part of the analysis requires unweighted data; using weighted data for the other criteria could produce spurious inconsistencies in the results.

To investigate the validity of inferences to the total population using unweighted analysis, we estimated classification errors from both weighted and unweighted data and observed that the classification error estimates expressed as proportions were virtually identical, differing only at the third decimal place. Thus, the results we report below using unweighted cell counts are appropriate for inference beyond the CPS sample to the total population.

Another consideration in using unweighted analysis is the estimation of standard errors. Since they are computed using simple random sampling assumptions, the ℓ EM standard error estimates may be understated as a result of ignoring the clustering effects in the CPS sample. To approximately account for this, we can multiply the ℓ EM variances by a design effect computed from the CPS labor force estimates. U.S. Bureau of the Census (2000, 14-9) indicates that the design effects for the CPS labor force estimates do not exceed 1.3 and thus multiplying the ℓ EM standard errors by $(1.3)^{1/2}$ should inflate the standard errors sufficiently to account for clustering. An equivalent approach is to use a 3 percent rather than a 5 percent level of significance in declaring the difference between two estimates to be statistically significant. This latter strategy will be employed in the forthcoming analysis as appropriate. We believe this produces a conservative test since the CPS design effect reflects the increase in variance due to both sample clustering and unequal weighting, while only clustering effects are present our unweighted estimates.

Table 1 shows the results of fitting a sequence of increasingly complex MLCA models for each of the three data sets. The Base Model is the simplest MLCA model and specifies that transition probabilities and response probabilities are homogeneous (*i.e.*, do not differ by group, P) and stationary (*i.e.*, are the same for all three months). This model may be written as

$$\pi_{p,a,b,c} = \sum_{x,y,z} \pi_p \pi_{x|g} \pi_{a|x}^3 \pi_{y|x}^2 \pi_{z|x} \quad (11)$$

which is obtained from (4) by imposing the constraints

$$\pi_{z|yp} = \pi_{y|xp} = \pi_{y|x} \quad (12)$$

and

$$\pi_{a|xp} = \pi_{b|yp} = \pi_{c|zp} = \pi_{a|x} \quad (13)$$

for all p .

For Model 1 we relax constraint (12) to

$$\pi_{z|yp} = \pi_{y|xp} \text{ for } p = 1, \dots, 4 \quad (14)$$

and thus allow transitions from January to February and February to March to vary by Self/Proxy Group, P . For Model 2, we further relax constraint (12) to

$$\pi_{y|xp} = \pi_{y|x} \text{ and } \pi_{z|yp} = \pi_{z|y} \quad (15)$$

Table 1
Model Diagnostics for Alternative MLCA Models by Year

1993 Data	<i>df</i>	<i>npar</i> ¹	L^2	<i>p</i> -value	BIC	<i>d</i>
Base Model: Homogeneous and stationary transitions and response probabilities	90	17	645	0	-320	0.048
Model 1: Nonhomogeneous transitions	84	23	632	0	-269	0.047
Model 2: Non-stationary transitions	66	41	99	0.006	-609	0.01
Model 3: Nonhomogeneous and non-stationary transitions	42	65	64	0.016	-386	0.01
Model 4: Nonhomogeneous and non-stationary transitions and nonhomogeneous response probabilities	24	83	23	0.501	-234	0
1995 Data	<i>df</i>	<i>npar</i> ¹	L^2	<i>p</i> -value	BIC	<i>d</i>
Base Model: Homogeneous and stationary transitions and response probabilities	90	17	697	0	-275	0.044
Model 1: Nonhomogeneous transitions	84	23	668	0	-240	0.043
Model 2: Non-stationary transitions	66	41	146	0	-567	0.01
Model 3: Nonhomogeneous and non-stationary transitions	42	65	82	0	-372	0.01
Model 4: Nonhomogeneous and non-stationary transitions and nonhomogeneous response probabilities	24	83	25	0.41	-234	0
1996 Data	<i>df</i>	<i>npar</i> ¹	L^2	<i>p</i> -value	BIC	<i>d</i>
Base Model: Homogeneous and stationary transitions and response probabilities	90	17	632	0	-325	0.045
Model 1: Nonhomogeneous transitions	84	23	585	0	-308	0.044
Model 2: Non-stationary transitions	66	41	159	0	-543	0.01
Model 3: Nonhomogeneous and non-stationary transitions	42	65	82.6	0	-364	0.01
Model 4: Nonhomogeneous and non-stationary transitions and nonhomogeneous response probabilities	24	83	39.3	0.026	-216	0

¹ Note that "npar" refers to the number of parameters in the model

for all p . Model 3 relaxes both the homogeneity and stationarity constraints for transition probabilities so that $\pi_{y|xp} \neq \pi_{z|yp}$. Thus this model allows transition probabilities to vary by group and by month. However, response probabilities are still constrained to be equal across groups and months.

Model 4 is the most general, identifiable model we considered. Model 4 allows the January-February and February-March transition probabilities to vary independently across the four proxy/self groups. This model further specifies that the response probabilities are the same for January, February, and March, but may vary across the four proxy/self groups. We obtained this model from Model 3 by relaxing the constraints specifying homogeneous response probabilities; i.e., by relaxing constraint (13) to

$$\pi_{a|xp} = \pi_{b|yp} = \pi_{c|zp} \quad (16)$$

for all p . Under these constraints, (4) can be written as

$$\pi_{p,a,b,c} = \sum_{x,y,z} \pi_p \pi_{x|p} \pi_{y|xp} \pi_{z|yp} (\pi_{a|p,x})^3.$$

In Table 1, we show the basic fit statistics for all five models for all three years. Column 4 of the table provides L^2 ,

the usual likelihood ratio chi-square statistic (see Agresti 1990, 48), and column 5 the corresponding p -value. A p -value of 0.05 or greater is the usual criterion for adequate model fit. However, due to the large sample sizes in our analysis, requiring a p -value this large could result in model over fitting. We consider a p -value as small as 0.01 to be acceptable. The BIC measure in the table is defined as

$$\text{BIC} = L^2 - (\log N)df$$

where N is the total sample size and df is the degrees of freedom for the model. The BIC essentially summarizes the tradeoff between model fit (L^2) and model parsimony (df). Since small values of the BIC are favorable, we will regard the model with the smallest BIC as best with respect to goodness of fit and parsimony. Liu and Dayton (1997) discuss this approach for latent class models.

Finally, the dissimilarity index (d) is the proportion of observations that would have to change cells for the model to fit perfectly. As rule of thumb, models having $d \leq 0.05$ (i.e., 5 percent model error) are considered to fit the data well (Vermunt 1997).

For each year of data, Model 4 is the only model to provide an acceptable fit when the p -value criterion is

Table 3
Comparison of MLCA Estimates with Prior Published Estimates

Classification		MLCA	Chua & Fuller (1982 data)	Poterba & Summers (1981 data)	CPS Reconciled Reinterview (1977-1982)
True	Observed				
Employed	Emp	98.77 (1993)	98.66 (month 1)	97.74	98.78
		98.73 (1995)	98.65 (month 2)		
		98.73 (1996)			
	Unemp	0.34 (1993)	0.32 (month 1)	0.54	0.19
		0.49 (1995)	0.34 (month 2)		
		0.37 (1996)			
	NLF	0.89 (1993)	1.02 (month 1)	1.72	1.03
		0.78 (1995)	1.01 (month 2)		
		0.79 (1996)			
Unemp	Emp	7.06 (1993)	3.52 (month 1)	3.78	1.91
		7.86 (1995)	3.51 (month 2)		
		8.57 (1996)			
	Unemp	81.81 (1993)	88.27 (month 1)	84.76	88.57
		76.09 (1995)	88.23 (month 2)		
		74.42 (1996)			
	NLF	11.13 (1993)	8.21 (month 1)	11.46	9.53
		16.04 (1995)	8.16 (month 2)		
		17.00 (1996)			
NLF	Emp	1.41 (1993)	1.60 (month 1)	1.16	0.5
		1.11 (1995)	1.61 (month 2)		
		1.13 (1996)			
	Unemp	0.75 (1993)	1.19 (month 1)	0.64	0.29
		0.69 (1995)	1.24 (month 2)		
		0.87 (1996)			
	NLF	97.84 (1993)	97.21 (month 1)	98.2	99.21
		98.20 (1995)	97.15 (month 2)		
		98.00 (1996)			

The table indicates that misclassification of the unemployed as NLF is a bigger problem than misclassification as Employed. Averaging over all three years, approximately two thirds of the error in classifying the unemployed is misclassification as NLF. But the rates of both types of error are high.

Next, we compare our estimates of the CPS classification probabilities with similar estimates from the literature. In Table 3, the MLCA estimates for each of the three years are compared with estimates from Chua and Fuller (1987), Poterba and Summers (1995), and the CPS reconciled reinterview program. Again, the latter three sets of estimates rely on reinterview data while the MLCA estimates are produced directly from the CPS interview data. In general, the relative magnitude of the MLCA estimates across the labor force categories agrees with the previous estimates. The greatest differences occur for the true unemployed population. For this group, the estimates of response accuracy from the literature are three to seven percentage points higher than corresponding MLCA estimates for 1993, which is the time period that most closely corresponds to the comparison estimates.

One explanation for this difference is that the comparison estimates are biased upward as a result of correlations between the errors in interview and reinterview. Another explanation is that the MLCA estimates are biased downward as a result of the failure of the Markov assumption to hold. We suspect that both explanations may be true to some extent. However, the next section provides some evidence that failure of the Markov assumption likely has a small effect on estimates of classification error.

3.6 Robustness of MLCA to Non-Markov Labor Force Transitions

A number of authors have investigated the effects of current and previous employment status on future employment status (see, for example, Akerlof and Main 1980; Heckman and Borjas 1980; Lynch 1989, and Corak 1993). Heckman and Borjas show that examination of this issue is quite difficult due to selection biases, response error, and unobserved heterogeneity. These confounding influences may account for the inconsistent findings in the literature. For example, using data from the CPS, Akerlof and Main (1980) provide evidence that the probability of

future unemployment depends upon the number of previous unemployment spells experienced as well as the duration of those spells. However, in a study of male high school graduates, Heckman and Borjas (1980) found "no evidence that previous occurrences of unemployment or their duration affect future labor market behavior once we control for sample selection bias and heterogeneity bias." The results from the literature are also inconsistent and ambiguous regarding the extent to which the Markov assumption expressed in (2) may be violated for the CPS and other labor market surveys. Nevertheless, in this section, we attempt to provide at least a partial answer to question of how non-Markov labor force transitions affect MLCA estimates of classification error.

To investigate the effect of violations of the Markov assumption in (2) for the present application, we conducted a limited simulation study. To focus the investigation while simplifying the simulation framework, we considered latent structures involving only two classes or states at each time point: unemployed, denoted by X, Y , or $Z = 1$, and other (*i.e.*, employed or not in the labor force), denoted by X, Y , or $Z = 2$ with analogous definitions for the observed states A, B , and C . To create a population for the simulation, the latent probabilities $\pi_{x^*}, \pi_{y|x^*}$, and $\pi_{z|x^*}$ and the response probabilities $\pi_{a|x^*} = \pi_{b|y^*} = \pi_{c|z^*}$ were specified to be consistent with the combined 1993, 1995, and 1996 data sets.

We then defined two parameters, λ_1 and λ_2 to be varied in the simulation, where

$$\lambda_1 = \frac{\pi_{z=1|x=2,y=1}}{\pi_{z=1|x=1,y=1}} \quad (17)$$

and

$$\lambda_2 = \frac{\pi_{z=1|x=2,y=2}}{\pi_{z=1|x=1,y=2}} \quad (18)$$

Thus, λ_1 is the probability of being "unemployed" in March, given "unemployed" in February and "other" in January over the probability of being "unemployed" in March given "unemployed" in the two previous months. Consistent with the findings of Akerlof and Main (1980) who showed that the likelihood of remaining unemployed increases as the number of unemployment spells increases, we assume that $0 \leq \lambda_1 \leq 1$. Similarly, λ_2 is the probability of being "unemployed" in March, given "other" in the two previous months, over the probability of being "unemployed", given "other" in February and "unemployed" in January. Again, by Akerlof and Main, we assume $0 \leq \lambda_2 \leq 1$. Note that when $\lambda_1 = \lambda_2 = 1$, unemployment transitions from February to March are Markov.

The simulated data were generated to be completely consistent with a MLCA model having non-stationary transition probabilities when $\lambda_1 = \lambda_2 = 1$. We simulated failure of the Markov assumption by varying λ_1 and λ_2 between 0 and 1. To be consistent with the 1993-1996 data, we fixed the probability of a correct "unemployed"

response, $\pi_{a=1|x=1}$, at 0.80 and the probability of a correct "other" response, $\pi_{a=2|x=2}$, at 0.99 in all simulations. In addition, the denominators of λ_1 and λ_2 were fixed to their values as determined from the combined 1993-1996 data while the numerators were computed from (17) and (18) using the values of λ_1 and λ_2 specified in each simulation run.

Table 4 summarizes the results of the simulation for $\lambda_1 = \lambda_2 = \lambda$ where λ is varied from 0.2 to 1.0 in steps of 0.2. Note that for $\lambda_1 = \lambda_2 = 1.0$, which corresponds to a Markov model, the estimated probabilities of correct response are exactly as specified. For smaller values of λ_1 and λ_2 , the estimates become negatively biased and are most biased for the lowest value considered, 0.2. Nevertheless, the absolute biases due to non-Markov transitions probabilities are never more than 3 percentage points. The results in Table 4 are consistent with Bushery and Kindelberger (1999), who used a somewhat different approach to illustrate the same robustness property of the MLCA models for CPS data. Both studies suggest that failure of the Markov assumption to hold does not appear to be an important source of bias in estimating CPS classification error probabilities.

Table 4
Estimates of Correct Classification Under
Non-Markov Transitions
(Cell entries are percentages)

Pr (Correct)	$\lambda_1 = \lambda_2 = \lambda$				
	$\lambda = 0.2$	$\lambda = 0.4$	$\lambda = 0.6$	$\lambda = 0.8$	$\lambda = 1.0$ (Markov)
Pr ("unemp" true "unemp") = $\pi_{a=1 x=1}$	77.6	78.1	78.7	79.3	80
Pr ("other" true "other") = $\pi_{a=2 x=2}$	98.6	98.7	98.8	98.9	99

4. COMPARING THE MLCA AND UNRECONCILED REINTERVIEW ESTIMATES

4.1 Hui-Walter Estimation

An alternative set of response probability estimates can be obtained from the CPS reinterview data using a type of latent class model first proposed by Hui and Walter (1980). Using the notation introduced above, let X denote the true labor force classification for some time point and let A and A' denote the interview and reinterview classifications, respectively. Let G denote a grouping variable defined as in (4). Consider the likelihood of the group \times interview \times reinterview table denoted by GAA' . Denote by $\pi_{gaa'}$ the probability of classifying an individual belonging to group g into cell (a, a') of the table. The model for $\pi_{gaa'}$ proposed by Hui and Walter is

$$\pi_{gaa'} = \sum_x \pi_g \pi_{x|g} \pi_{a|x} \pi_{a'|x} \quad (19)$$

NOMINATIONS SOUGHT FOR THE WAKSBERG INVITED PAPER SERIES

Survey Methodology has established an annual invited paper series in honor of Joseph Waksberg, who has made many important contributions to survey methodology. Each year, as part of the Waksberg Invited Paper Series, a prominent survey researcher will be chosen to author a paper that will review the development and current state of a significant topic within the field of survey methodology, and will reflect the mixture of theory and practice that characterizes Waksberg's work. The author will receive a cash award made possible by a grant from Westat, in recognition of Joe Waksberg's contributions during his many years of association with Westat. The grant is administered financially by the American Statistical Association.

The author will be selected by a four-person committee appointed by *Survey Methodology* and the American Statistical Association. Nominations of individuals to be considered as authors should be sent to the chair of the committee, Graham Kalton, at Westat, 1650 Research Blvd., Rockville, MD 20850, by e-mail to kalton1@westat.com, or by fax to 301-294-2034. Nominations for the author of the 2002 Waksberg paper must be received by March 16, 2001.

NOMINATIONS RECHERCHÉES POUR LA SÉRIE D'ARTICLES SOLLICITÉS WAKSBERG

La publication *Techniques d'enquête* comprend maintenant une série annuelle d'articles sollicités en l'honneur de Joseph Waksberg, dont la contribution au domaine de la méthodologie d'enquête a été importante. Chaque année, dans le cadre de la série d'articles sollicités Waksberg, un éminent chercheur du domaine des enquêtes sera choisi pour rédiger un article dans lequel il examinera l'évolution et l'état actuel d'un sujet significatif en méthodologie d'enquête. Cet article reflètera la combinaison pratico-théorique typique des travaux de Waksberg. Westat versera à l'auteur un prix en argent, en reconnaissance de l'apport de Joe Waksberg pendant ses nombreuses années de service à Westat. Les fonds de ce prix sont gérés par l'American Statistical Association.

L'auteur sera sélectionné par un comité formé de quatre personnes nommées par *Techniques d'enquête* et l'American Statistical Association. Le nom des auteurs proposés doit être envoyé à Graham Kalton, président du comité, Westat, 1650, boul. Research, Rockville, Maryland 20850, par courriel à kalongl@westat.com, ou par télécopieur au (301) 294-2034. Le nom des auteurs proposés pour l'article Waksberg de 2002 doit parvenir d'ici le 16 mars 2001.

In this model, the parallel measures assumption for the interview and reinterview responses is relaxed and response probabilities for the two measures, *viz.* $\pi_{a|x}$ and $\pi_{a'|x}$, are estimated separately. The ICE assumption is made as a condition of identifiability. It is further assumed that $\pi_{a|x}$ and $\pi_{a'|x}$ do not depend upon the group variable, G , while the prevalence of employed, unemployed, and NLF, *i.e.* $\pi_{x|g}$, still depends upon G .

Sinclair and Gastwirth's (1996) analysis of CPS labor force classification error used Sex as the grouping variable and our analysis uses this grouping variable as well. Sinclair and Gastwirth confined their analysis to white males and females and two labor force categories: NLF and In the Labor Force. The latter category is the sum of our Employed and Unemployed categories. In our analysis, we consider sample members of all races and analyze the three category labor force classification used in the MLCA. Thus, the H-W analysis estimates 16 parameters for each year, which equals the number of degrees of freedom available from the $G \times A \times A'$ table, leaving no degrees of freedom to test model fit.

The ℓ EM software was used to fit the H-W model to the interview and unreconciled reinterview data from three time periods that coincide with the three in our MLCA: pre-1994, 1995, and 1996. We attempted to restrict the analysis to only the first quarter of these time periods. Unfortunately due the small sample sizes, the estimates were quite unstable. Thus, it was necessary to use the reinterview data from all four quarters of these time periods. The pre-1994 data were collected from 1985 through 1988 using the unreconciled reinterview sample.

The results of this comparison of MLCA and H-W estimates are summarized in Table 5. The MLCA estimates are the same as those in the rows of Table 2 labeled "Total." The H-W estimates are the classification probabilities associated with the original interview, *i.e.*, measure A in (19). The table shows the comparison for all three years. Since the largest error rate in the MLCA occurred for the Unemployed, this category is of particular interest in the MLCA/H-W comparison.

Overall, the two sets of estimates show fairly good agreement. The years 1995 and 1996 exhibit no statistically significant differences (at the 5 percent level) between the MLCA and H-W estimates for the unemployed population. The pre-1994 estimates display significant differences; however, they may be explained by the fact that the pre-1994 reinterview data were from 1985 through 1988, rather than 1993. These differences will be explored further in the next section.

4.2 Comparison of Indexes of Inconsistency

As described in section 3.1, we compute estimates of the index of inconsistency for each time period using the MLCA model-based estimates of the response probabilities. Essentially, we estimate the expected interview-reinterview cross-classification table from the MLCA response

probability estimates and then apply the formula for the index to this table as though the table were observed. A second expected interview-reinterview classification table can be estimated using the H-W response probability estimates. We then compared these two sets of estimates to the estimate of the index computed directly from the CPS reinterview data using traditional methods (U.S. Bureau of the Census 1985). Agreement of all the three estimates agree supports the validity of the three methods.

Table 5
Comparison of MLCA and H-W Model Estimates of CPS
Response Probabilities by Year
(Standards Errors are in Parentheses)

Classification		1993		1995		1996	
True	Observed	H-W	MLCA	H-W	MLCA	H-W	MLCA
Emp	Emp	99.3 (0.3)	98.8 (0.1)	99.5 (0.7)	98.7 (0.1)	99.6 (0.1)	98.8 (0.1)
	Unemp	0.0 (0.0)	0.3 (0.1)	0.0 (n/a)	0.5 (0.1)	0.4 (0.1)	0.4 (0.1)
	NLF	0.7 (0.3)	0.9 (0.1)	0.5 (0.7)	0.8 (0.1)	0.0 (n/a)	0.8 (0.1)
Unemp	Emp	11.1 (1.0)	7.1 (0.7)	11.5 (2.3)	7.9 (0.9)	4.6 (15.2)	8.6 (1.0)
	Unemp	74.3 (2.7)	81.8 (1.1)	67.9 (6.1)	76.1 (1.3)	67.6 (11.1)	74.4 (1.4)
	NLF	14.7 (2.9)	11.1 (0.9)	20.6 (6.5)	16.0 (1.2)	27.9 (5.3)	17.0 (1.2)
NLF	Emp	2.0 (0.5)	1.4 (0.1)	2.5 (1.5)	1.1 (0.1)	2.6 (1.5)	1.1 (0.1)
	Unemp	1.2 (0.3)	0.8 (0.1)	0.5 (0.6)	0.7 (0.1)	0.0 (n/a)	0.9 (0.1)
	NLF	96.8 (0.6)	97.8 (0.1)	97.0 (1.6)	98.2 (0.1)	97.4 (1.1)	98.0 (0.1)

Table 6 shows the three methods estimates the index of inconsistency for all three time periods. As before, the Unemployed category is of particular interest because of its large error rate. Standard errors are not available for the MLCA or the H-W estimates of the index so formal tests of hypothesis are not possible. However, standard errors for the traditional estimates are provided which can be used as rough approximations of the standard errors for the H-W estimates.

Overall, both the general patterns of the MLCA estimates and the magnitudes of the MLCA estimates generally agree quite well with the H-W and traditional estimates for all three years. However, for the NLF category in 1995 and 1996, the traditional estimates of I are somewhat larger than either of the latent class model estimates. Further analysis suggests that this difference is due to a bias in the traditional estimation approach resulting from the failure of the parallel measures assumption.

U.S. Bureau of the Census (1985) shows that if the interview and reinterview processes have different reliabilities, then the traditional estimate of the index will be biased. For example, if the reliability of the reinterview

data is lower than the reliability of the interview data, the traditional test-retest reliability estimator will understate the actual reliability of the CPS data; *i.e.*, the CPS index of inconsistency will be too large.

Table 6

Comparison of MLCA, H-W, and Traditional Estimates of the Index of Inconsistency by Year and Labor Force Classification

Method of Estimation	Labor Force Classification			Aggregate Index
	Employed	Unemployed	Not in Labor Force	
1993				
Traditional estimation	8.16 (0.24)	33.49 (1.16)	9.96 (0.27)	11.05 (0.26)
H-W	7.37	34.93	10.07	10.78
MLCA	6.35	28.04	7.63	8.73
1995				
Traditional estimation	6.69 (0.44)	36.28 (2.85)	10.80 (0.56)	10.42 (0.53)
H-W	6.82	37	8.98	9.7
MLCA	6.06	36.19	7.2	8.72
1996				
Traditional estimation	5.93 (0.39)	35.97 (2.68)	11.95 (0.56)	10.61 (0.51)
H-W	5.67	39.46	7.55	8.56
MLCA	5.99	37.39	7.76	9.06

The CPS interview and reinterview will have different reliabilities if the error distributions for the two interviews are not equal. A test of this is possible by comparing the fit of a H-W type model with and without the restriction $\pi_{a|x} = \pi_{a'|x}$. The assumption of equal reliability is rejected if the difference between the likelihood ratio chi-squares for the two models exceeds a chi-square with 6 degrees of freedom. This test was rejected for 1995 and 1996 at the 10 percent level of significance. Thus, it appears that the difference in the NLF estimates for 1995 and 1996 may be due, in part, to bias in the traditional estimates of I .

Note further that the H-W and MLCA indexes agree quite well for 1995 and 1996, although they differ somewhat in 1993. However, as noted in the discussion of Table 5, the comparisons between the MLCA and H-W estimates for this year are confounded by the different time periods used to construct the pre-1994 interview-reinterview data set. This could account for at least some of the discrepancy between the estimates for this year.

5. SUMMARY AND CONCLUSIONS

The primary goal of this research was to investigate the validity of MLCA estimates of CPS labor force classification error and to determine the efficacy of MLCA as an alternative to traditional methods for evaluating CPS data quality. We analyzed interview data from the CPS for the

first quarter of three years – 1993, 1995, and 1996 – and conducted an additional analysis of the CPS unreconciled reinterview data for approximately the same time periods. The reinterview data provided another approach for estimating CPS classification error that, when compared with the MLCA estimates, helped to address the question of the validity of the MLCA approach.

Five dimensions of MLCA validity were addressed as follows:

1. **Model diagnostics.** We investigated a wide range of MLCA models with grouping variables defined by age, race, sex, education, mode of interview, and proxy/self response. The most parsimonious and best fitting model for all three years included one grouping variable defined by the proxy/self variable with four categories: all three waves conducted by self response, only two waves conducted by some self response, only two waves conducted by proxy response, and all three waves conducted by proxy response. For this class of models, the best model was Model 4 (see Table 1) which specified non-homogeneous and non-stationary transition probabilities and non-homogeneous response probabilities. This model provided an adequate fit to the data for all three years.
2. **Model Goodness of Fit Across Years of CPS.** Another indicator of model validity is its fit across independent samples of the same population. Assuming that labor force dynamics and the response probability structure for the CPS is stable across the span of four years, the same general model should fit all three years adequately. Model 4 displays multi-year goodness of fit (see Table 1). In addition, other grouping variables were tested in the study, yet the proxy/self variable model emerged as the best variable for all three years.
3. **Agreement Between the Model and Test-Retest Estimates of Response Probabilities.** Using the unreconciled interview-reinterview data from the CPS for the time periods pre-1994, 1995, and 1996, we applied the H-W method to estimate the response probabilities and compared these with the MLCA estimates. There was good agreement for 1995 and 1996, the two years for which the time periods for the reinterview data and the CPS data were closely matched (see Table 5). For 1993, we observed small but significant differences between MLCA estimates and the corresponding H-W estimates. These differences might be explained by differences in the time periods, since the reinterview data predated the CPS interview data by some years.
4. **Agreement Between the Model and Test-Retest Estimates of Inconsistency.** We compared MLCA model-based estimates of the index of inconsistency with the corresponding direct estimates from the CPS

reinterview program. The two sets of estimates agree fairly well for all three years, with the exception of the NLF category (see Table 6). For 1995 and 1996, the differences can be partly explained by the bias in the traditional estimator resulting from the failure of the parallel measures assumption. The H-W method, which does not require the assumption of parallel measures, produces estimates of the index that agree well with MLCA estimates for 1995 and 1996. For 1993, the difference between the MLCA and H-W estimates may be due to the difference in the time periods for the reinterview and the CPS data sets.

5. Plausibility of the Patterns of Classification Error.

The MLCA estimates of misclassification probabilities appear to be plausible. The estimates across proxy/self groups were consistent with prior expectations that lower error rates should be observed for self respondents than for proxy respondents. In addition, the largest error rates were observed for the unemployed population and the magnitudes of these estimates were consistent with those of previous studies – for e.g., Fuller and Chua 1985; Abowd and Zellner 1985; Porterba and Summers 1986; and Sinclair and Gastwirth 1996 (see Table 3).

In summary, we found no evidence from these analyses to question the validity of the MLCA approach. The method performed well in all five validity tests. We therefore recommend that the MLCA method be considered as an alternative method for evaluating the accuracy of the CPS labor force estimates. The strong agreement between the MLCA and H-W estimates supports the validity of the H-W method as well. We recommend that both methodologies be considered in future studies of CPS data quality.

Although the MLCA approach performed well in our tests, we recommend caution in applying the methodology in other settings. In our analysis, reinterview data provided a means for assessing the validity of the MLCA estimates. However, reinterview data are typically not available in panel surveys and, consequently, analysts may only be able to apply criteria (1), (2), and (5) above to check model validity. The Markov assumption is key to the MLCA approach. Some panel data may seriously violate this assumption. Fortunately, failure of Markov assumption appears not to be an important factor in the validity of MLCA estimates of CPS labor force classification error (*cf.* Table 4).

REFERENCES

- ABOWD, J., and ZELLNER, A. (1985). Estimating gross labor-force flows. *Journal of Business and Economic Statistics*, 3, 3, 254-283.
- AGRESTI, A. (1990). *Categorical Data Analysis*. New York: John Wiley & Sons.
- AKERLOF, G.A., and MAIN, G.M. (1980). Unemployment spells and unemployment experience. *The American Economic Review*, 70, 3, 885-893.
- BAILAR, B. A. (1975). The effect of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-30.
- BIEMER, P., BUSHERY, J., and FLANAGAN, P. (1997). An Application of Latent Markov Models to the CPS. Internal U.S. Bureau of the Census Technical Report.
- BIEMER, P., and FORSMAN, G. (1992). On the quality of reinterview data with applications to the Current Population Survey. *Journal of the American Statistical Association*, 87, 420, 915-923.
- BOHRNSTEDT, G.W. (1983). Measurement. *Handbook of Survey Research*, (P.H. Rossi, R.A. Wright, and A.B. Anderson, Eds.). New York: Academic Press.
- BUSHERY, J., and KINDELBERGER, K. (1999). Simulation Examples for MLC Analysis. Internal U.S. Bureau of the Census Memorandum, Washington, DC, 70-122.
- CHUA, T.C., and FULLER, W.A. (1987). A model for multinomial response error applied to labor flows. *Journal of the American Statistical Association*, 82, 397, 46-51.
- CLOGG, C., and ELIASON, S. (1985). Some common problems in log-linear analysis. *Sociological Methods and Research*, 16, 8-14.
- COHEN, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 210, 37-46.
- CORAK, M. (1993). Is unemployment insurance addictive? Evidence from the benefit durations of repeat users. *Industrial and Labor Relations Review*, 47, 1, 62-72.
- FORSMAN, G., and SCHREINER, I. (1991). The design and analysis of reinterview: an overview. *Measurement Errors in Surveys*, (P.P. Biemer, et al., Eds.). New York: John Wiley & Sons. 279-302.
- FULLER, W., and CHUA, T.C. (1985). Gross change estimation in the presence of response error. *Proceedings of the Conference on Gross Flows in Labor Force Statistics*. Washington, D.C., U.S. Bureau of the Census and U.S. Bureau of Labor Statistics, 65-77.
- HECKMAN, J.J., and BORJAS, G.J. (1980). Does unemployment cause future unemployment? Definitions, questions, and answers from a continuous time model of heterogeneous and state dependence. *Economica*, 47, 247-283.
- HESS, J., SINGER, E., and BUSHERY, J. (2000). Predicting test-retest reliability from behavior coding. *International Journal of Public Opinion Research*, II, 4, 346-360.
- HUI, S.L., and WALTER, S.D. (1980). Estimating the error rates of diagnostic tests. *Biometrics*, 36, 167-171.
- KLEINBAUM, D.G., KUPPER, L.L., and MULLER, K.E. (1988). *Applied Regression Analysis and Other Multivariate Methods*. Boston: PWS-KENT Publishing Co.
- LIU, T.H., and DAYTON, C.M. (1997). Model selection information criteria for non-nested latent class models. *Journal of Educational and Behavioral Statistics*, 22, 249 - 264.

- LYNCH, L.M. (1989). The youth labor market in the eighties: determinants of re-employment probabilities for young men and women. *The Review of Economics and Statistics*, 37-45.
- MEYERS, B. D. (1988). Classification-error models and labor-market dynamics. *Journal of Business and Economic Statistics*, 6, 3, 385-390.
- MOORE, J.C. (1988). Self/proxy response status and survey response quality. *Journal of Official Statistics*, 4, 2, 155-122.
- O'MUIRCHARTAIGH, C. (1991). Simple response Variance: Estimation and Determinants. *Measurement Errors in Surveys*, (P. Biemer *et al.*, Eds.). New York: John Wiley & Sons, 551-574.
- POTERBA, J., and SUMMERS, L. (1986). Reporting errors and labor market dynamics. *Econometrics*, 54, 6, 1319-1338.
- POTERBA, J., and SUMMERS, L. (1995). Unemployment benefits and labor market transitions: a multinomial logit model with errors in classification. *The Review of Economics and Statistics*, 77, 207-216.
- POULSEN, C.S. (1982). Latent Structure Analysis with Choice Modeling Applications. Doctoral dissertation, Wharton School, University of Pennsylvania.
- ROTHGEB, J. (1994). Revisions to the CPS Questionnaire: Effects on Data Quality, U.S. Bureau of the Census. CPS Overlap Analysis Team Technical Report 2, April 6.
- SCHREINER, I. (1980). Reinterview Results from the CPS Independent Reconciliation Experiment (Second Quarter 1978 through Third Quarter 1979). Internal U.S. Bureau of the Census Report.
- SHOCKEY, J. (1988). Adjusting for response error in panel surveys, a latent class approach. *Sociological Methods and Research*, 17, 1, 65-92.
- SINCLAIR, M., and GASTWIRTH, J. (1996). On procedures for evaluating the effectiveness of reinterview survey methods: application to labor force data. *Journal of the American Statistical Association*, 91, 961-969.
- SINCLAIR, M., and GASTWIRTH, J. (1998). Estimates of the errors in classification in the labour force survey and their effects on the reported unemployment rate. *Survey Methodology*, 24, 2, 157-169.
- SINGH, A.C., and RAO, J.N.K. (1995). On the adjustment of gross flow estimates for classification error with application to data from the canadian labour force survey. *Journal of the American Statistical Association*, 90, 430, 478-488.
- U.S. BUREAU OF THE CENSUS (1985). Evaluating Censuses of Population and Housing, STD-ISP-TR-5. Washington, D.C.: U.S. Government Printing Office.
- U.S. BUREAU OF THE CENSUS (2000). Current Population Survey: Design and Methodology. U.S. Bureau of the Census Technical Paper 63, Washington, D.C.: Government Printing Office.
- VAN DE POL, F., and DE LEEUW, J. (1986). A latent markov model to correct for measurement error. *Sociological Methods and Research*, 15, 1-2, 118-141.
- VAN DE POL, F., and LANGEHEINE, R. (1997). Separating change and measurement error in panel surveys with an application to labor market data. *Survey Measurement and Process Quality*, (L. Lyberg, *et al.*, Eds.). New York: John Wiley & Sons.
- VERMUNT, J. (1997). *4EM: A General Program for the Analysis of Categorical Data*. Tilburg University.
- WIGGINS, L.M. (1973). *Panel Analysis, Latent Probability Models for Attitude and Behavior Processing*, Amsterdam: Elsevier S.P.C.

Estimation and Replicate Variance Estimation of Median Sales Prices of Sold Houses

KATHERINE J. THOMPSON and RICHARD S. SIGMAN¹

ABSTRACT

The U.S. Census Bureau publishes estimates of medians for several characteristics of new houses, with a key estimate being sales price of sold houses. These estimates are calculated from data acquired from interviews of home builders by the Survey of Construction (SOC). The SOC is a multi-stage probability survey whose sample design is well suited to the modified half-sample replication (MHS) method of variance estimation. The literature supports applying the MHS method to replicate sample medians to estimate the sampling variance of a median. There are several computational advantages, however, to using grouped data to estimate medians, with linear interpolation being used within the grouped-data interval containing the median. Using survey data and simulated finite populations, we compared the effects of no grouping (*i.e.*, the sample median), grouping with fixed-size intervals, and grouping with data-dependent-sized intervals on medians and associated MHS variance estimates. We examined the mean squared errors and mean absolute errors of the median estimates and the relative bias and stability of the variance estimates and the coverage of the associated confidence intervals. We found that the data-dependent-sized intervals yielded variance estimates with the smallest bias, the best stability, and the best confidence intervals.

KEY WORDS: Median; Modified half-sample replication; Survey of Construction.

1. INTRODUCTION

The U.S. Census Bureau publishes estimates of medians for several characteristics of new houses, with a key estimate being sales price of sold houses. These estimates are calculated from data acquired from interviews of home builders by the Survey of Construction (SOC). The SOC is a multi-stage probability survey whose sample design is well suited to the modified half-sample (MHS) replication method (balanced repeated replication with replicate weights of 1.5 and 0.5) for reasons outlined in section 3.B. In the near future, the SOC will move its current estimation and variance estimation systems to the Census Bureau's re-engineered post-data-collection system, the Standardized Economic Processing System (StEPS). When this occurs, SOC will change from its current non-replicate variance estimation procedure to the MHS replication variance estimation procedure (Thompson 1998). Because the SOC variance estimation methodology is changing, we decided to revisit the median-estimation methodology for continuous data. Our goal was to find a median-estimation method with good estimation and variance estimation properties, given the MHS replication.

We considered two methods of median-estimation. The first method uses the sample weights to estimate medians via empirical cumulative-distribution functions. The second method uses linear interpolation of grouped continuous data to approximate the median. The latter method is implemented in VPLX (Variances from ComPLex Survey, Fay 1995), the replicate variance estimation software package developed at the Census Bureau.

Direct calculation of sample medians can be computationally intensive because it requires separate sorts for each

value of a given classification variable. An alternative estimation method is to group the continuous data into discrete intervals (called bins) and use linear interpolation over the interval containing the median. Provided that the data are approximately uniformly distributed over the interval containing the median, interpolation yields a good approximation while being considerably less computer resource-demanding. However, optimal bin widths and locations may differ by domain and may change over time as the sample distributions change.

In this paper, we compare six methods of median-estimation, given MHS replication: the sample median and five variations using linear interpolation. Section 2 provides a brief overview of the SOC design. Section 3 presents general methodology. Section 4 describes the empirical results from four months of SOC data that motivated the simulation study presented in section 5. Section 6 provides our conclusions and recommendations.

2. SOC SAMPLE DESIGN

The SOC universe contains two sub-populations: local areas that require building permits and local areas that do not. The SOC sample-units selected from the first sub-population comprise the Survey of the Use of Permits (SUP), and those selected from the second sub-population, the Nonpermit Survey (NP). The SUP sample comprises the majority of the SOC estimate. The two samples are multi-stage probability samples stratified by variables with high expected correlation with the survey's key statistics: housing starts, completions, and sales.

¹ Katherine J. Thompson and Richard S. Sigman, Economic Statistical Methods and Programming Division, U.S. Census Bureau, Washington DC, 20233, U.S.A.

The first stage of the SUP and NP sample selection is a subsample of 1980 design Current Population Survey (CPS) Primary Sampling Units (PSUs), which are contiguous areas of land with well-defined boundaries. Thus, both surveys are conducted in the same PSUs but are otherwise independent samples. The PSUs were stratified within region by weighted 1980 population 16 years and older, weighted 1982 residential permit activity, and percent of housing in nonpermit areas. When possible, strata consisted of PSUs from the same state with the same metropolitan status. One PSU per stratum was selected. Self-representing (SR) PSUs were included in the sample with certainty (the stratum consists of one PSU). Nonself-representing (NSR) PSUs were selected with probability proportional to size (PPS) from strata containing more than one PSU.

The second stage of SUP sample selection is a stratified systematic sample of permit-issuing places within sample PSUs (selected once a decade). These places were stratified by a weighted average of the ratio of permit-issuing activity in year j to the total US permit activity in year j ($j = 78, 81, 82$). In many cases, only one second stage unit was selected. The third stage of SUP sample selection is performed monthly: each month, Field Representatives (FRs) select a systematic sample of building permits from the permit offices in each sampled permit-issuing place. One-to-four-unit building permits are selected systematically in such a way that an overall one-in-forty sample is achieved; five-or-more-unit building permits are included with certainty. The third-stage samples are independent by month; the first and second stages are not.

The second stage of NP sample selection is a stratified systematic sample of small land areas (1980 Census Enumeration Districts, or EDs), stratified by 1980 Census population size. For the third stage of NP sample selection, field representatives completely canvass all of the roads in the sampled EDs (called segments). To reduce canvassing, a few of the larger EDs were subsegmented and one subsegment selected, or large EDs were 1-in-2 subsampled. Currently, there are a total of seventy-one active nonpermit segments. All new housing units are included in the NP sample with certainty.

Median estimates are derived from the pooled SUP and NP samples and are calculated using a post-stratified weight for the SUP portion and an unbiased weight for the NP portion.

3. METHODOLOGY

A. Median-Estimation Procedures

1. Sample Median

One procedure for estimating the median of a population is to calculate the sample median from ungrouped data, using the sample weight to locate the median. This approach is recommended in Kovar, Rao and Wu

(1988) and Rao and Shao (1996). The procedure uses the following steps:

- sort the sample observations in ascending order;
- accumulate the sum of the associated survey weights;
- select the first observation for which the associated sum of the weights exceeds fifty percent of the total weight.

2. Linear Interpolation

Another approach for estimating the median of a population is to group the sample data and interpolate for the sample median. Woodruff (1952) provides the following formula for linear interpolation of a sample median:

$$\hat{M} = F^{-1}\left(\frac{1}{2}\hat{N}\right) \approx ll + \left(\frac{\frac{1}{2}\hat{N} - cf}{f_i}\right) * (i) \quad (3.1)$$

where

F = the cumulative frequency of the characteristic using sample weights

ll = lower limit of the bin containing the median

\hat{N} = estimated total number of elements in the population

cf = cumulative frequency in all intervals preceding the bin containing the median

f_i = median class frequency (estimated total number of elements in the population of the interval containing the median)

i = width of the bin containing the median

This is the method used by the current SOC production variance estimation system for monthly estimates and is also the linear interpolation method employed by VPLX.

We considered two options for setting the class size (bin widths) for the interpolation. The first option develops bins based on the specific characteristic under consideration using the original data. The second option linearly transforms the data to a standard scale and then uses a standard set of bins for every characteristic. We used the following linear transformation:

$$X'_i = X_i * \frac{1,000}{Q_3} \quad (3.2)$$

where Q_3 is the third quartile of the sample distribution (estimated using the ordered observations and sample weight as outlined in section 2.A.1). The interpolated median of the X' is multiplied by $(Q_3/1,000)$ to obtain an estimated median on the original scale [If the distribution contains negative values (e.g., a distribution of net income), then use $X'_i = (X_i - X_{(1)}) * 1,000/Q_3(X_i - X_{(1)})$, where $X_{(1)}$ is

the first order statistic and $Q_3(X_i - X_{(1)})$ is calculated from the distribution of $(X_i - X_{(1)})$. To obtain an estimated median on the original scale, multiply the interpolated median by $(Q_3(X_i - X_{(1)})/1,000)$ and add $X_{(1)}$.] This procedure is equivalent to simply dividing the original sample from 0 to Q_3 into \underline{x} bins of equal width and placing the remainder of the data into one bin which, by design, is much larger than the others.

This procedure is designed for symmetric or positively skewed distributions (usually the case with economic data). The data in the last bin is not used to estimate the median because it is greater than Q_3 , which is expected to be far from the median. If we based the linear transformation on Q_1 (the first quartile), the bin containing the median might be very close to the lowest bin in the distribution. In this case, the difference in variability between an interpolated median and the sample median would be small.

Using the original data to develop medians has the advantage of producing production-ready estimates and SEs. Determining the appropriate fixed bin width is difficult, however. As the bin widths get small (approach width 1), the variance estimates become more unstable. As the bin widths increase, the bias of the estimate due to interpolation increases. The "optimal" bin size balances variance estimate stability and bias. Unfortunately, the optimal bin width may not remain constant between samples. Often, the distributions change over time, and the bins widths/locations in the sample should reflect this change in scale. Moreover, the optimal bin width may be different for different values of a classification variable: for example, the optimal bin width for the Midwest's sales price is probably different from the optimal bin width for the South's sales price.

The desire to have the width of the bin depend on the sample motivated the linear transformation. The "standard" bin widths used for the transformed data less than Q_3 are not standard on the untransformed scale: the bin width is data-dependent. Using the linearly transformed data requires more bookkeeping in terms of scaling constants but easily allows for changes in the scale and shape of the distribution.

Figures 1 through 4 illustrate the effect of the linear transformation on the bin widths and location for two distributions. Figures 1 and 2 present a distribution that has a large spread of data values, including a few very large observations. Figures 3 and 4 present a distribution consisting of primarily small data values.

Figure 1 presents a histogram of the original distribution for houses sold with conventional financing, with bin width of \$25,000 [Note: the bin size was selected purely for presentation convenience, since this is a long-tailed distribution]. The median of this distribution is \$167,130, and Q_3 is \$225,000. Figure 2 presents the histogram of the linearly transformed distribution with bin width of 50. In this example, the transformed bins of width 50 are equivalent to bins of width \$11,250 on the original scale

((\$225,000/1,000)*50). Recall that the original-data bin sizes considered are \$1,000 and \$2,000. Thus, the transformed-data bins of width 4 would have a width of \$900 on the original untransformed scale. Notice the large "spike" at the last bin, which contains all of the sample greater than Q_3 .

These figures also illustrate the differences in distribution of sample sizes across bins between the two methods. Using fixed bin widths with the original data results in quite variable bin sample sizes (see Figure 1). In contrast, by design the sample sizes within the data-dependent bins are much more uniform for all but the last bin (see Figure 2).

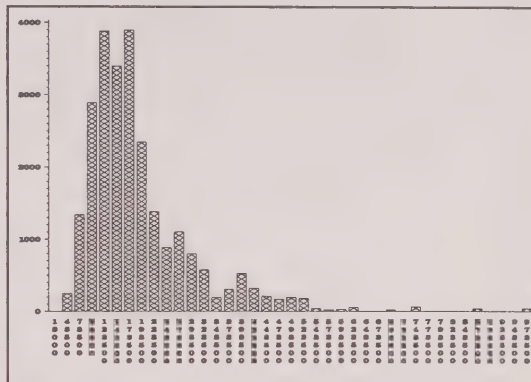


Figure 1: Original Distribution of Sales Price of Houses Sold With Conventional Financing Bin Width = \$25,000

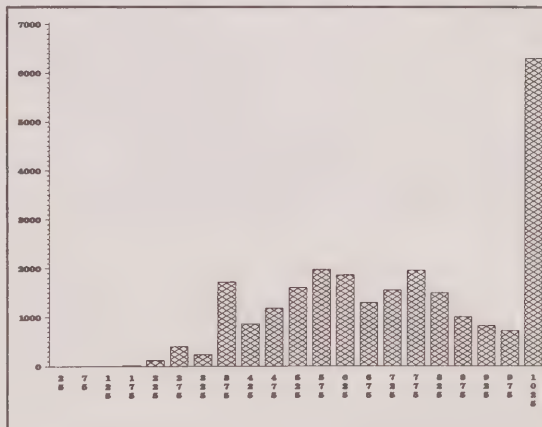


Figure 2: Transformed Distribution of Sales Price of Houses Sold With Conventional Financing Using Bin Width = 50 Bin Width on Untransformed Scale = \$11,250

Figure 3 presents a histogram of the original distribution of houses sold with FHA loans, with bin width of \$4,000 (again, the bin width is chosen for presentation convenience). The median of this distribution is \$108,280, and Q_3 is \$124,990. Figure 4 presents the histogram of the linearly transformed distribution with bin width of 50. In this example, the transformed bins of width 50 are

equivalent to bins of width \$6,250 on the original scale, and the transformed-data bins of width 4 would have approximate width \$500 on the original untransformed scale.

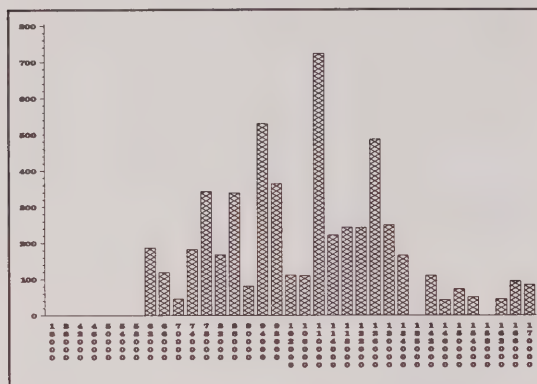


Figure 3: Original Distribution of Sales Price of Houses Sold With FHA Loans Bin Width = \$4,000

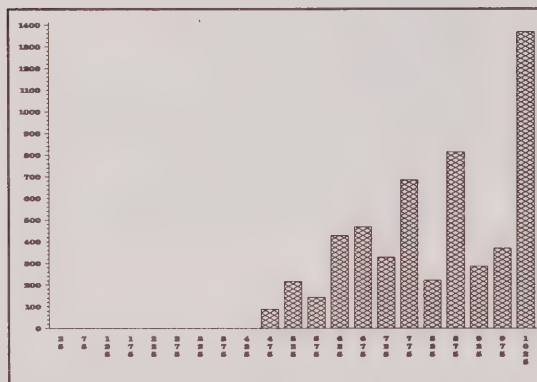


Figure 4: Transformed Distribution of Sales Price of Houses Sold With FHA Loans Using Bin Width = 50 Bin Width on Untransformed Scale = \$6,250

Figures 1 through 4 demonstrate the flexibility of the bins developed for linearly-transformed data. The bin size on the untransformed scale expands or contracts, depending on the spread of the data. Moreover, the data-dependent bin sample sizes are less variable compared to those associated with fixed bins.

To evaluate the first interpolation option (original-data-interpolated medians), we used two different sets of bin widths (classification sizes): bins of size \$2,000 (the same bin width used in the current production variance estimation system) and bins of size \$1,000. [Note: The VPLX variance estimation software would not allow any bin size smaller than 1,000 because the number of classes exceeded the allowable array range.] After examining several months of sales price estimates for the total U.S., we assumed that median sales price would always be larger than \$36,000 and smaller than \$550,000, so the first original-data classification is always (low – 35,999) and the last original-data

classification is always (550,000 – high): this yields 257 bins of size \$2,000 or 514 bins of size \$1,000, plus one bin of size \$36,000 and one bin whose width depends on the largest observation in the sample. One obvious problem with the locations of these bins is the potential effect of inflation. It is conceivable that within special financing categories or certain regions, the median sales price for houses sold could approach \$550,000, and the interpolation would fail as a consequence.

To evaluate the second interpolation option (transformed-data-interpolated-medians), we used three different sets of bin widths: bins of size 4, 25, and 50. The bins of size 4 were chosen to be analogous to the bins of size 2,000 in terms of the number of bins. There are 250 bins of size 4 for the transformed data less than Q_3 , and one larger bin containing all data greater than Q_3 . The selection of widths 25 and 50 was somewhat arbitrary: we chose bin size 50 to get a total of twenty bins for the data less than Q_3 ; and we chose bin size 25 to examine the effect of doubling the number of bins/halving the width of the bins for data less than Q_3 . The transformed-data median is always less than 1,000, so the last transformed-data classification is always (1,000 – high). Thus, by definition the last bin contains up to twenty-five percent of the data and is considerably wider than the other bins.

B. Variance Estimation

We used the Modified Half-Sample (MHS) replication method (Fay 1989 and Judkins 1990) to estimate the variance of a median as supported in the literature (e.g., Rao, Wu, and Yue (1992); Rao and Shao (1996); Kovacevic and Yung (1997) for balanced repeated replication; and Judkins (1990) for MHS replication). MHS replication is a variation of the “traditional” balanced half-sample variance estimation described in Wolter (1985, 110-152). Balanced half-sample replication (BRR) is a variance estimation method designed for a two-PSU per stratum design. With BRR, a half-sample replicate is formed by selecting one unit from each pair and weighting the selected unit by 2 (so that it represents both units). Thus, estimates for every PSU are included in each replicate although half are weighted by zero. Replicates (half-samples) are specified using a Hadamard matrix. See Wolter (1985, 114-115) for a detailed description of the replicate formation procedure using Hadamard matrices. MHS replication uses replicate weights of 1.5 and 0.5 in place of the 2 and 0. The standard error for a median estimate using MHS replication is given by

$$\hat{SE}(\hat{Med}) = \sqrt{\frac{4}{R} * \sum_{r=1}^R (\hat{Med}_r - \hat{Med}_0)^2}$$

where the r subscript refers to the replicate r median estimate ($r=1, 2, \dots, R$) and the 0 subscript refers to the full sample the median estimate. This expression contains a four (4) in the numerator because the MSE of the replicate

estimates is too small by a factor of $1/(1-0.5)^2$. See Judkins (1990).

Neither the SUP nor the NP designs are two-sample-unit-per-stratum designs. At the first stage, one PSU per stratum is selected. The second and third stages are systematic samples, and often only one unit per stratum was selected at the second stage. A common approach used to address the one sample-unit per stratum problem is to

- “split” the SR sample-units into two panels per sample-unit using the original sampling methodology;
- form collapsed strata by pairing two (or three) “similar” NSR sample-units; and
- apply the half-sample approach in such a way that the elements contributing to the half samples are panels within sample-units for SR sample-units and are the first stage sample-units (PSUs) within collapsed strata for NSR sample-units.

The current SOC production variance system uses a Keyfitz estimator (a paired difference estimator) for NSR sample and an approximate sampling-formula estimator for SR sample to produce level estimate variances (Luery 1990). Because SOC methodologists had already collapsed NSR strata for their paired difference estimator, a BRR-like application was a logical extension of the pre-existing variance estimation structure. For MHS replication, we sort permits within predetermined sample-unit groups in SR units by geography and authorization date and systematically split the ordered sample into two panels as suggested in Wolter (1985, 131). Although this is essentially the only approach available for the SOC design, this method may not provide the correct variance estimates since units in both panels are correlated (in the original half-sample method, the two PSUs in the stratum are assumed independent). For more details on the replicate assignments, see Thompson (1998).

The SOC production system uses the Woodruff method (Woodruff 1952) to estimate the standard error of a median. The Woodruff method uses the estimated SE of a proportion \hat{p} ($\hat{p} = 0.50$ for median-estimation) and projects the interval ($\hat{p} \pm \text{SE}(\hat{p})$) through the cumulative frequency distribution to obtain the lower limit of a 62.86 percent confidence interval for the median (the $\text{SE}(\hat{p})$ can be estimated using replicate methods). The SE of the median is then estimated by subtraction. This methodology has had mixed success in the past according to SOC survey analysts.

4. EMPIRICAL DATA RESULTS

Initially, we used four months of SOC sample data to examine the variances of the median-estimation methods for sales price of sold houses: March 1997, May 1997, June 1997, and July 1997. We produced medians by region and by type of financing. We used the same weight used by the

SOC production estimation and variance systems (post-stratified for SUP sample and unbiased for NP sample), pooling both surveys' data to obtain medians. Each set of variance estimates was produced using 200 replicates.

We found that the six median-estimation methods produced very similar estimates, but yielded three distinct sets of SEs: one set for the sample median, one set for the original-data-interpolated medians (fixed bin width), and one set for the transformed-data-interpolated medians (data-dependent bin width). There was no clear relationship between bin width and SE estimates within the two sets of interpolated medians. Indeed, within type of data (original or transformed), the SEs were all very close. Clearly, there was a linear transformation and an interpolation effect. None of the median-estimation methods yielded standard errors resembling the published standard errors, so there was no available argument for publication consistency.

Moreover, there is some evidence that the Woodruff method publication SEs are underestimates or are at least inappropriate for the sample design used. Kovar, Rao, and Wu (1988) compared Woodruff SEs and BRR standard errors and found that the two methods had similar properties except for the case of stratified samples, where the strata are based on highly correlated separate variables (such as the SOC design). In this case, the Woodruff SE is often too small, and they concluded that “the BRR... methods (sic) are more robust to different population structures, since the error is extracted directly from the replicates.” When the production system Woodruff SEs used the directly-calculated $\text{SE}(\hat{p})$, the Woodruff SEs were generally smaller than the replicate SEs.

The empirical results left us in a quandary. We had three distinct sets of variance estimates, and no “gold standard” against which to measure them. Because our empirical results were inconclusive, we conducted a Monte Carlo simulation study to evaluate the properties of the MHS variance estimates produced from the different median estimators.

5. SIMULATION STUDY COMPARISON

A. Procedure for Simulation Study

We created four finite artificial populations based on a data analysis of four SOC sample populations: one type-of-financing population (Conventional Financing) and three regional populations (Midwest (Region 2), South (Region 3), and West (Region 4)). These populations represented a variety of the types of SOC populations from which estimates are produced. Note that the SOC type-of-financing population is not independent of the SOC-region populations.

To approximate the finite population of sales price for houses sold, we generated w_i records for each sample-unit i , where w_i is the sample weight associated with unit i . The distributions of sales price for single-unit sold houses could

be approximated by lognormal distributions. The lognormal distribution has the probability density function

$$f(y) = \frac{1}{y - \theta} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2} \left(\frac{(\log(y - \theta) - \zeta)}{\sigma}\right)^2\right)$$

for $\theta < y < \infty$

where θ is the threshold parameter, ζ is the scale parameter, and σ is the shape parameter.

From our models, we generated four simulated finite bivariate populations with expected correlation $\rho = 0.6$ using the method outlined in Naylor, Balintfy, Burdick and Chu (1968, 99). The first of the two variables in each population represented sales price of sold houses and was obtained by generating a random normal variable with mean ζ and variance σ^2 using the parameters determined above, then exponentiating and shifting by the appropriate location parameters (θ). The second variable was used to form strata and first stage clusters. This variable had a marginal standard normal distribution and was obtained by independently generating a second standard random normal value, multiplying it by 0.8, and adding this term to $0.6 \times$ the standard normal random variable used to generate the sales price variable. Percentiles, sample skewness, and sample kurtosis of each simulated population's sales price variable were very close to the corresponding statistics in the original population, especially when outliers were deleted using the resistant outer fences rule described in Hoaglin and Iglewicz (1987). Each population's size was the \hat{N} estimated from the sample populations. Model parameters, sample correlations (between simulated sales price and stratifying variable), population size (N), and sample sizes (n) are reported in Table 1.

After generating the finite populations, we formed 50 equal sized strata in each population, then selected two sets of samples for two different survey designs:

- The first design is patterned after the SUP sample of permits for four-or-less-housing units in SR permit offices in SR PSUs (approximately 28% of the SOC sample). In this study, we selected 5,000 stratified without-replacement random samples from each simulated population using the same sampling rate in each stratum. To perform MHS replication, we sorted the sample within each stratum by stratifying variable and then systematically split the sample into two panels.

- The second design is patterned after the SUP sample of permits for four-or-less-housing units in NSR permit offices in SR PSUs and in SR permit offices in NSR PSUs (approximately 40% of the SOC sample). In this study, we selected 5,000 two-stage samples from each simulated population. The first stage is stratified without-replacement random sample of two PSUs per stratum ($N_h = 5$). The second stage is a systematic sample of units within PSUs. Because all PSUs are the same size, this study does not take the SOC PPS sampling into account and does not include the collapsing of first-stage units. The MHS replication uses the first-stage sample units (PSUs) within the same strata. The replicate weights do not account for large sampling fractions at the first stage of selection as recommended in Wolter (1985, 122), so all of the variance estimates are probably upwardly biased.

We did not attempt to simulate the SUP sample of permits for four-or-less-housing units in NSR PSUs and NSR permit offices (a three-stage sample, approximately 25% of the SOC sample); the SUP sample of permits for five-or-more housing units (approximately 2% of the SOC sample); or the NP sample of EDs (approximately 5% of the SOC sample). The three-stage sample, although non-negligible in SOC, is rarely used by other surveys at the Census Bureau, and the other two sectors of the SOC design do not contribute enough to the estimates to warrant a separate investigation.

To examine the precision of each median-estimation procedure over repeated samples, we estimated empirical Mean Squared Errors (MSE) and Mean Absolute Errors (MAE) from the 5,000 samples for:

- SM:** the sample median of each half-sample
- IO2000:** interpolated medians using original data, bins of size 2,000 (fixed bin width)
- IO1000:** interpolated medians using original data, bins of size 1,000 (fixed bin width)
- IT4:** interpolated medians using linearly transformed data, bins of size 4 (data dependent bin width)
- IT25:** interpolated medians using linearly transformed data, bins of size 25 (data dependent bin width)
- IT50:** interpolated medians using linearly transformed data, bins of size 50 (data dependent bin width)

Table 1
Characteristics of Simulated Populations and Sample Sizes of Stratified Samples

Population	Distribution	Sales Price Parameters			Correlation (Stratifier, Sales Price)	Population Size	Sample Size
		θ	σ	ζ	ρ	N	n
Conventional Financing	lognormal	27,578	0.4895	11.84	0.57030	25,150	500
Midwest	lognormal	31,801	0.5957	11.69	0.55835	6,500	150
South	lognormal	29,414	0.5549	11.55	0.55929	14,550	300
West	lognormal	53,781	0.5822	11.59	0.55525	11,550	250

Table 2
Median, Third Quartile, and Bin Widths on Original Scale for Transformed Simulated Data

Population	Median	Q_3	Bin Width		
			4	25	50
Conventional Financing	167,173	222,263	889	5,557	11,113
Midwest (Region 2)	151,312	210,647	843	5,266	10,532
South (Region 3)	133,745	180,868	723	4,522	9,043
West (Region 4)	162,130	214,320	857	5,358	10,716

The linear transformation was performed once for procedures IT4, IT25, and IT50. The original data were transformed using the full sample Q_3 , and these transformed data were assigned to the half-samples (including replicate 0, the full sample). Table 2 provides the median and third quartile of each finite population, along with the bin widths on the original scale for the transformed data.

We calculated $M(\zeta_i)$, the empirical MSE of median-estimation procedure i as

$$\begin{aligned} M(\zeta_i) &= \frac{\sum_r (\zeta_{ri} - \bar{\zeta}_i)^2}{5,000} + (\bar{\zeta}_i - \zeta_p)^2 \\ &= \hat{\sigma}^2(\zeta_i) + \text{bias}^2(\zeta_i) \end{aligned} \tag{5.1}$$

where ζ_{ri} is the estimated median for sample r and estimator i , $\bar{\zeta}_i$ is the average of the ζ_{ri} , and ζ_p is the population median. This is the empirical MSE described in Judkins (1990).

We calculated the Mean Absolute Error (MAE) of each median-estimation procedure i as

$$MAE(\zeta_i) = \frac{\sum_r |\zeta_{ri} - \zeta_p|}{5,000} \tag{5.2}$$

as defined in DeGroot (1986, 209-211).

To compare the variance estimation properties of the different median-estimation methods, we calculated an MHS variance estimate (v_{ij}) corresponding to each median-estimation procedure i from 1,000 of the 5,000 samples. These variance estimates were compared in terms of

Relative bias $(\sum_{j=1} v_{ij}/1,000)/M(\zeta_i) - 1$

Relative stability $[(\sum_{j=1} (v_{ij} - M(\zeta_i))^2/1,000)]^{1/2}/M(\zeta_i)$

Error Rate Number of samples where $(\zeta_p < \theta_{Li} \text{ or } \zeta_p > \theta_{Ui})/1,000$ where

θ_{Li} is the lower end of a 90% confidence interval, and

θ_{Ui} is the upper end of a 90% confidence interval

These criteria are used in Kovar, Rao, and Wu (1988) and in Rao and Shao (1996). The relative bias is a measure of the bias of the variance estimate as a proportion of the true MSE. The stability is a measure of the variance of the variance estimates; it approximates a c.v. of the variance estimate v_i . Note that the relative stability is not the relative MSE defined in Wolter (1985, 297) which uses the squared-MSE in the denominator. With an “optimal” variance estimator, both the relative bias and relative stability will be near zero, and the error rate will be ten percent.

B. Results

1. Comparison of Median-estimation Procedures

Table 3 presents the empirical root MSE, standard error, the bias, and the MAE for each median-estimation procedure from both simulation studies. Each of these statistics was calculated from 5,000 samples.

These results reinforced our suspicions from the empirical data analysis described earlier. At least for sales price, all six median-estimation procedures perform approximately equally well, with approximately equal root-MSEs and MAEs between procedures in each population.

2. Comparison of MHS Replication Variance Estimation Properties of Median-Estimation Procedures

When we examined the variance estimation properties for each procedure, the results were quite different. As with our empirical data analysis, we had three very distinctive sets of results. Table 4 summarizes the three different comparison measures for the variance estimates in the four populations. The numerators for the relative bias and stability and the coverage rates are based on 1,000 samples. The denominator for the relative bias and stability (“truth”) are based on 5,000 samples. An asterisk (*) in the last column of Table 4 indicates that the error rate is significantly different from the nominal error rate of 0.10 using the normal approximation to the binomial distribution at the 90% confidence level.

Table 3
Empirical Root MSE, Standard Error, Bias, and MAE for Median-Estimation Procedures

Population	Median-Estimation Procedure	Unclustered Single-Stage Sample				Clustered Two-Stage Sample			
		Root MSE	SE	Bias	MAE	Root MSE	SE	Bias	MAE
Conventional Financing	SM	3,345	3,345	-12	2,671	3,389	3,374	324	2,733
	IO2000	3,320	3,316	161	2,698	3,346	3,341	189	2,685
	IO1000	3,387	3,368	-354	2,642	3,431	3,420	-278	2,774
	IT4	3,351	3,340	273	2,673	3,378	3,364	311	2,719
	IT25	3,304	3,293	276	2,617	3,337	3,321	322	2,664
Region 2 Midwest	IT50	3,282	3,265	329	2,606	3,305	3,283	375	2,636
	SM	6,316	6,287	-598	4,966	6,273	6,228	-753	4,959
	IO2000	6,276	6,275	-127	4,992	6,335	6,207	-1,271	5,029
	IO1000	6,343	6,297	-767	4,939	6,526	6,280	-1,774	5,204
	IT4	6,372	6,363	328	5,004	6,294	6,228	-908	4,979
Region 3 South	IT25	6,273	6,272	127	4,937	6,270	6,154	-1,199	4,971
	IT50	6,220	6,218	160	4,936	6,224	6,114	-1,164	4,966
	SM	3,670	3,658	301	2,931	3,835	3,752	796	3,054
	IO2000	3,708	3,669	539	2,998	3,796	3,739	656	3,011
	IO1000	3,742	3,740	101	2,941	3,809	3,804	212	3,066
Region 4 West	IT4	3,718	3,662	639	2,951	3,814	3,736	766	3,028
	IT25	3,699	3,638	669	2,924	3,793	3,711	787	2,992
	IT50	3,692	3,616	745	2,912	3,778	3,680	856	2,970
	SM	4,385	4,382	-140	3,509	4,394	4,351	616	3,506
	IO2000	4,425	4,421	185	3,578	4,362	4,339	449	3,487
	IO1000	4,477	4,469	-258	3,530	4,411	4,410	-57	3,535
	IT4	4,414	4,403	318	3,514	4,383	4,342	599	3,494
	IT25	4,376	4,364	315	3,460	4,334	4,296	573	3,439
	IT50	4,367	4,350	391	3,455	4,320	4,271	644	3,436

In both studies, the variance estimates of the transformed-data-interpolated medians perform best in terms of relative bias and stability. Specifically,

- The variance estimates of the transformed-data-interpolated medians (IT4, IT25, IT50) have the smallest relative bias. The difference in estimation method is quite pronounced in three of the four populations, where the largest relative bias of the transformed-data-interpolated medians is less than one-half the size of the smallest relative bias of the original-data-interpolated and sample medians. These results are surprisingly strong for the two-stage clustered design, since the variance estimates are expected to be biased upwards (see section 5.A);
- The variance estimates of the interpolated medians had the best stability. The variance estimates of the sample median had the poorest stability in all four populations. This result was expected due to the smoothing effect of interpolation. Again, the transformed-data-interpolated medians generally performed better than the original-data-interpolated medians, although the difference is not as pronounced as in the case of relative bias. Generally, the stability is close with all three bin widths for the transformed-data-interpolated medians.

The results for each median-estimation procedure’s confidence interval coverage are not as consistent, varying by design. With the single-stage unclustered design, the

confidence intervals constructed from transformed-data-interpolated medians and SEs have the best coverage. In each population, the data-dependent bins (all widths) yield close to nominal or better coverage; in fact, none of these error rates is statistically different from the nominal 10%. The confidence intervals constructed from original-data-interpolated medians and SEs are extremely conservative. Here, the positive bias in the variance estimates makes these intervals unnecessarily wide, thereby reducing the power to make interesting findings. The coverage with the sample median is erratic.

Some of these coverage patterns are repeated in the two-stage clustered design. Again, the coverage with the sample median is erratic, and the coverage rates for the confidence intervals constructed from original-data-interpolated medians are better than nominal (although only significantly better than nominal in two populations). The error rate pattern is quite different for the transformed-data-interpolated medians. In all but the Region 4 population, the coverages rates for the three procedures are worse than nominal. However, with bins of widths 4 and 25, only one error rate is significantly larger than 10%; for bins of width 50, two of these three error rates are significantly larger than 10%. All of the interpolated-data-medians have significantly smaller than nominal error rates in the Region 4 population; consistent with the other population’s results, the error rates for the original-data-interpolated medians are the farthest from 10%.

Table 4
Relative Bias and Relative Stability for Variance Estimates, and Error Rates for 90% Confidence Intervals

Population	Median-Estimation Procedure	Unclustered Single Stage Design			Clustered Two-Stage Design		
		Relative Bias	Relative Stability	Error Rate	Relative Bias	Relative Stability	Error Rate
Conventional Financing	SM	0.19	0.69	11.0%	0.11	0.58	15.1%*
	IO2000	0.25	0.35	6.9%*	0.25	0.37	9.0%
	IO1000	0.21	0.32	7.0%*	0.19	0.33	9.3%
	IT4	0.06	0.25	10.0%	0.06	0.27	11.3%
	IT25	0.07	0.25	10.9%	0.06	0.27	11.8%*
	IT50	0.05	0.26	9.5%	0.05	0.28	12.1%*
Region 2 Midwest	SM	0.57	1.24	7.3%*	0.41	1.07	7.9%*
	IO2000	0.33	0.44	6.9%*	0.23	0.35	8.6%
	IO1000	0.30	0.42	7.0%*	0.17	0.30	8.7%
	IT4	0.15	0.41	10.1%	0.14	0.41	11.5%*
	IT25	0.16	0.40	9.8%	0.11	0.37	10.4%
	IT50	0.15	0.42	9.0%	0.11	0.40	10.4%
Region 3 South	SM	0.30	0.88	12.4%*	0.15	0.71	11.1%
	IO2000	0.31	0.42	6.7%*	0.28	0.39	7.5%*
	IO1000	0.29	0.40	6.7%*	0.27	0.38	7.3%*
	IT4	0.04	0.29	11.0%	0.01	0.28	10.8%
	IT25	0.02	0.28	11.0%	-0.01	0.27	11.3%
	IT50	0.01	0.29	11.1%	-0.02	0.28	11.9%*
Region 4 West	SM	0.39	0.98	8.9%	0.25	0.79	8.6%
	IO2000	0.32	0.42	6.2%*	0.31	0.41	5.2%*
	IO1000	0.29	0.39	6.2%*	0.28	0.38	5.2%*
	IT4	0.11	0.32	8.6%	0.10	0.31	7.6%*
	IT25	0.10	0.31	9.4%	0.09	0.30	7.5%*
	IT50	0.08	0.31	9.5%	0.08	0.31	8.3%*

In both studies, the transformed-data-interpolated medians have the best variance estimation properties in terms of relative bias and relative stability by a large margin, regardless of bin width. And, in both studies, the transformed-data-interpolated medians using bins of width 4 or width 25 have excellent confidence interval coverage. Since the transformed-data-interpolated-medians using bins of width 50 or width 25 yielded the “best” estimators in terms of root-MSE and MAE in both studies, using linear interpolation on transformed data with bins of width 25 appears to be the best median-estimation procedure in terms of estimation and variance estimation properties.

6. CONCLUSION

We explored the effect of using variations of two different methods of estimating the median sales price of sold houses: direct estimation versus linear interpolation. Linear interpolation requires classifying continuous data into bins of standard width. This width can be arbitrary, can differ greatly by domain, and may change as the sample distribution changes over time. The linear transformation

based on the third quartile appeared to correct this problem. With the transformed data, the bins’ widths and locations in the distribution change depending on the data.

Our empirical results indicated that the choice of method has a pronounced impact on the variance estimates given MHS replication. Our simulation study examined the properties of the different median-estimation procedures on the MHS replicate variance estimates. In all four simulated populations, the transformed-data-interpolated medians (data dependent bin widths) performed the best, usually by a wide margin. Most critically, this method greatly reduces the overestimation of the variance. Using bins of width 25 on the transformed scale (41 bins total) yielded the best median sales price estimates and variance estimates, given MHS replication and is our recommended method for the Survey of Construction.

The recommended method has several advantages. First, it is adaptive. It works well for a variety of distributions, because the bin widths themselves depend on the distribution at hand. Second, it saves computing resources by avoiding sorting half-samples. Third, the data-dependent-intervals can be easily incorporated into generalized survey-processing software. Finally, it gives better estimates and

MHS replicate variance estimates (at least for sales price of sold houses). We expect that these results are generalizable for other continuous distributions as well, although obviously this conjecture should be tested on other data sets. Other areas for future research include examining the relationship between sample size and precision of the median estimates, examining alternative bin sizes, and exploring the robustness of the recommended procedure with different replicate variance estimation procedures.

ACKNOWLEDGEMENTS

The authors would like to thank Elizabeth Huang and James Fagan of the U.S. Census Bureau, two anonymous referees, and the associate editor for their helpful comments on earlier versions of this manuscript, and J.N.K Rao for his useful comments on the original simulation study. This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.

REFERENCES

- DeGROOT, M. (1986). *Probability and Statistics*. Reading, MA: Addison-Wesley Publishing, Inc.
- FAY, R.E. (1989). Theory and application of replicate weighting for variance calculations. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.
- FAY, R.E. (1995). VPLX: Variance Estimation for Complex Surveys. Program Documentation: Unpublished Bureau of the Census Report.
- HOAGLIN, D.C., and IGLEWICZ, B. (1987). Fine-tuning some resistant rules for outlier labeling. *Journal of the American Statistical Association*, 83, 1147-1149.
- JUDKINS, D.R. (1990). Fay's method for variance estimation. *Journal of Official Statistics*, 6, 223-239.
- KOVAR, J.G., RAO, J.N.K., and WU, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *The Canadian Journal of Statistics*, 16, 25-45.
- KOVACEVIC, M., and YUNG, W. (1997). Variance estimation for measures of income inequality and polarization – An empirical Study. *Survey Methodology*, 23, 41-52.
- LUERY, D.M. (1990). Survey of Construction Technical Paper. Unpublished draft Bureau of the Census internal documentation.
- NAYLOR, T.H., BALINTFY, J.L., BURDICK, D. S., and CHU, K. (1968). *Computer Simulation Techniques*. New York: John Wiley and Sons, Inc.
- RAO, J.N.K., WU, C.F.J., and YUE, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, 18, 209-217.
- RAO, J.N.K., and SHAO, J. (1996). On balanced half-sample variance estimation in stratified random sampling. *Journal of the American Statistical Association*, 91, 343-348.
- THOMPSON, K.J. (1998). Evaluation of Modified Half-Sample Replication for Estimating Variances for the Survey of Construction (SOC). Technical Report #ESM-9801, available from the Economic Statistical Methods and Programming Division.
- WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag, Inc.
- WOODRUFF, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 635-646.

The Impact of Different Rotation Patterns on the Sampling Variance of Seasonally Adjusted and Trend Estimates

C.H. McLAREN and D.G. STEEL¹

ABSTRACT

Many economic and social time series are based on sample surveys which have complex sample designs. The sample design affects the properties of the time series. In particular, the overlap of the sample from period to period affects the variability of the time series of survey estimates, and the seasonally adjusted and trend estimates produced from them. The Census X11 and X11ARIMA packages are commonly used to produce seasonally adjusted estimates and can also be used to produce estimates of trend. This paper considers the implications of different overlap patterns on the sampling variance of seasonally adjusted and trend estimates obtained from time series based on sample surveys.

KEY WORDS: X11; X11ARIMA; Seasonal adjustment; Trend estimation; Rotation patterns.

1. INTRODUCTION

Many important time series are based on repeated sample surveys which have complex patterns of sample overlap from period to period. The use of sampling means that the estimated time series have a component of variability due to sampling errors and for many series this will be a major source of variability. The sample design, in particular the overlap pattern, affects the variability of the time series of survey estimates.

Increasingly, analysis of time series is concentrating on assessing underlying patterns of change or trends based on analysis of the seasonally adjusted series. Most government statistical agencies have calculated seasonally adjusted series for many years. Kenny and Durbin (1982) noted that policy analysts frequently say that they are more interested in underlying trends than following irregular fluctuations in the de-seasonalized monthly values. A similar view is expressed by Smith (1997). For more than 10 years the Australian Bureau of Statistics (ABS) has published series of trend estimates obtained by applying Henderson Moving Averages (HMAs) (Henderson 1916) to the seasonally adjusted series to smooth out the irregular components of the series (ABS 1987). Other government statistical agencies also produce trend estimates using a variety of methods (Knowles 1997). Since seasonally adjusted and trend estimates are obtained by processes applied to the original series, they are also influenced by sampling errors. Bell and Kramer (1999) note that the variance of seasonally adjusted estimates will often be dominated by the contribution from sampling error. Some series are based on independent samples over time, but usually the samples used have a degree of overlap from period to period to reduce costs and the standard errors of estimates of change between two consecutive time periods (Kish 1998).

A key issue in the development of the design of a repeated survey is the rotation pattern, that is, the pattern of a selected unit's inclusion in the survey over time, which will determine the sample overlap. The aim of this paper is to determine the effects of the rotation pattern used on the sampling variance of the estimated seasonally adjusted and trend series obtained using the Census X11 method developed by Shiskin, Young and Musgrave (1967) and X11ARIMA developed by Dagum (1980 and 1988). We will focus on the estimates of the level and one period change in the seasonally adjusted and trend estimates.

2. ROTATION PATTERNS

Consider a univariate time series with values y_t , $t = 1, \dots, T$, obtained from a repeated sample survey. The observed value at time t is related to the true value of the series in the finite population, Y_t , by

$$y_t = Y_t + e_t$$

where e_t is the sampling error. The series Y_t is thought to consist of trend-cycle, seasonal and irregular components T_t , S_t and I_t , so that

$$y_t = T_t + S_t + I_t + e_t.$$

In some cases a multiplicative decomposition may be more appropriate. Many statistical agencies produce seasonally adjusted series by attempting to estimate S_t and remove it from the series, usually using some combination of linear filters. Most commonly used is the Census X11 method developed by Shiskin *et al.* (1967) and X11ARIMA developed by Dagum (1980 and 1988). Findley, Monsell, Otto, Bell and Pugh (1998) described further enhancements embodied in X12ARIMA. The ABS also publishes trend

¹ C.H. McLaren and D.G. Steel, School of Mathematics and Applied Statistics, University of Wollongong, NSW 2522, Australia. E-mail: craigmcl@uow.edu.au, dsteel@uow.edu.au.

estimates obtained by applying HMAs to the seasonally adjusted series and encourages users to base their interpretation of the series on these trend estimates (Linacre and Zarb 1991; ABS 1993). The HMAs were originally derived by Henderson (1916) for use in actuarial work and are used within X11, X11ARIMA and X12ARIMA to de-trend series for seasonal adjustment purposes. Kenny and Durbin (1982) and Gray and Thomson (1996) explain the derivation of the HMAs. Users can also produce trend estimates by applying filters to the published seasonally adjusted estimates. Kenny and Durbin (1982) noted that there is no unique definition of trend and that different filters may be used according to the degree of smoothness and sensitivity required. Knowles and Kenny (1997) investigated methods of trend estimation for official statistical series. For monthly series they recommended the use of HMAs, with the length of the filter being 13 or 23 depending on the volatility of the series in question.

The autocorrelation structure of the observed series is determined by the autocorrelation of the series Y_t and e_t , which will then affect the estimates of the trend, seasonally and irregular components. The covariance structure of the sampling error series, e_t , can be estimated from the unit level survey data. By obtaining such estimates, it is possible to obtain estimates of the sampling variance of the estimated trend, seasonally adjusted and irregular series. Various methods for doing this have been proposed; for example Steel and DeMel (1988) considered the effect of linear filters on the spectrum of the sampling error series and Wolter and Monsour (1981) used an approach based on the effect of linear filters on the autocovariance function. Sutcliffe (1993) adopted a similar approach using a linear approximation to the X11 procedure. Pfeiffermann (1994) proposed a method which develops an estimate of sampling error directly from the estimated time series using various simplifying assumptions. These approaches do not explicitly model the time series. Other authors, for example Bell and Wilcox (1993), Tiller (1992), Burridge and Wallis (1985) and Hausman and Watson (1985), considered explicit ARIMA models for both the true series and the sampling error series, and concentrated on the estimation of the parameters of the models. These papers do not consider the effect of different rotation patterns and concentrate on producing estimates of the variances of seasonally adjusted estimates for the particular rotation pattern used.

The rotation pattern used in the survey will affect the autocorrelation structure of the sampling error series and hence the sampling variance of the original, seasonally adjusted and trend estimates. Several considerations are taken into account in deciding upon a rotation pattern. High sample overlap between consecutive periods reduces the sampling variance of estimates of change between the periods and high sample overlap between periods 12 months apart reduces the sampling variance of estimates of annual change. The first occasion that a selected unit is included in the survey is usually the most expensive. By keeping selected units in the survey for longer the cost of the survey is

reduced. This leads to rotation patterns in which a selected unit is included every period for as long as possible. However, a selected unit must eventually be rotated out of the survey. Besides the ethical consideration of spreading respondent load, there is the possible deterioration in response rate and quality of data reported if the same unit is included for a large number of occasions (see Kalton and Citro 1993, for a discussion of these issues).

Rotation patterns vary in terms of the number of times a unit is included in the survey and the time interval between inclusions. We concentrate on monthly labour force surveys (MLFSs). The rotation patterns used in practice are special cases of the a - b - $a(m)$ rotation patterns where selected units are included for a consecutive months, removed from the survey for b months then re-included for a further a months. The pattern is repeated so that selected units are included for a total of m occasions. Rao and Graham (1964) considered the estimation of the finite population means and totals for this class of rotation patterns. The United States Current Population Survey (CPS) uses a 4-8-4(8) pattern (Fuller, Adam and Yansaneh 1992). Putting $b = 0$ gives an *in-for- m* rotation pattern in which selected dwellings are included for m months after which they are removed from the sample. The case $m = 6$ corresponds to the Canadian rotation pattern (Singh, Drew, Gambino and Mayda 1990) and $m = 8$ corresponds to the Australian pattern (ABS 1992). Steel (1997) noted that the British quarterly labour force survey approximately corresponds to a monthly survey with a 1-2-1(5) rotation pattern.

We consider the sampling variance of the seasonally adjusted and trend estimates associated with the rotation patterns currently used in MLFSs and a number of rotation patterns that, while not currently used, may have some desirable properties. This will give an indication of which rotation patterns are better in terms of the component of the variability of the estimated series that is affected by the sample design.

3. SAMPLING VARIANCE OF SEASONALLY ADJUSTED AND TREND ESTIMATES

Let y_T be the vector containing the values of the time series of survey estimates up to time T and Y_T be the vector containing the true population values. The sampling variance of the original series is denoted by $V(y_T|Y_T)$. Consider a linear filter which is used to obtain values from y_T by applying a vector of filter weights w_t . The filter weights are non-random and have no connection with the survey weights used in calculating the survey estimates y_t . The filter weight vectors w_t depends on the time period for which the filtered value refers. The weights are constant within the body of the series but may be modified at the beginning and end. The filtered value at time t is

$$\tilde{y}_t = w_t' y_T. \quad (1)$$

Then

$$V(\tilde{y}_t | Y_T) = w'_t V(y_T | Y_T) w_t \quad (2)$$

is the sampling variance of the filtered value at time t . The sampling error of the filtered value is the difference between $w'_t y_T$ and $w'_t Y_T$, which is conditional on the values of the true series, Y_T . This is the difference between the filtered value obtained from the series of estimates ending at time T and the value that would be obtained if that series was observed without sampling error. We focus on this component as it is the sampling variance that can be altered by changing the sample design. The variance associated with Y_T has not been taken into account. Wolter and Monsour (1981) discussed the issue of total variance versus sampling error variance. There may be advantages in considering the total variance in interpreting the resulting series but when we are considering sample design issues, such as the choice of rotation pattern, we focus on the component that is directly affected by decisions made about the sample design. If the sampling error does not contribute significantly to the variability of the series then decisions about the sample design are not as important as they are when the sampling error is a major contributor, although it still seems sensible to use as effective a sample design as possible.

To determine the effect of different rotation patterns on the sampling variance of a particular filtered series, we need an estimate of $V(y_T | Y_T)$ for different rotation patterns. Previous work on estimating variances of seasonally adjusted series has either ignored the rotation pattern and assumed independent samples at each time point, or taken it as fixed and used an estimation method that takes it into account. We need a model for $V(y_T | Y_T)$ that reflects the effect of the different rotation patterns that could be used.

The analysis of the effect of different rotation patterns is simplified if the series of sampling errors has a stable autocorrelation structure. The precise form of the autocorrelation function will depend on the series and should reflect the complexities of the design. For example Steel and DeMel (1988) suggested a model for the Australian Monthly Labour Force data and Bell and Wilcox (1993) suggested a model for the United States Retail Trade series. Bell and Hillmer (1990) and Miazaki and Dorea (1993) also considered modelling of survey errors by time series models. Dempster and Hwang (1993) and Lee (1990) considered approaches to estimating and modelling sampling error correlations for the US CPS.

Our approach is to assume that the series of sampling errors, e_t , has constant variance. A model is needed for the correlation between the sampling errors of y_t and y_{t+s} . All the rotation patterns considered imply that the sample at any particular time will consist of a number of panels. A panel is a set of units that are included and removed from the survey at the same time. When a panel is rotated out of the survey it will be replaced by another panel. The set of panels related in this way is referred to as a rotation group. Most MLFSs use multistage sampling and when a panel is rotated out of

the survey it is replaced by another panel of nearby households (see ABS 1992; Singh *et al.* 1990). Hence it is assumed that the sampling correlation between estimates obtained from the same rotation group s periods apart is $r(s)$ if no rotation has occurred and $d(s)$ if rotation has occurred. We will assume that the estimate at time t is, at least approximately, the average of estimates from each rotation group and that estimates from different rotation groups, which will usually be in different PSUs and spatially well separated, are independent.

These assumptions imply that the sampling correlation between y_t and y_{t+s} is

$$R(s) = d(s) + k(s)(r(s) - d(s)) \quad (3)$$

where $k(s)$ is proportion of the sample in common between the two time periods. The sample overlap factor $k(s)$ is determined by the rotation pattern. For example, for an *in-for-m* rotation pattern $k(s) = 1 - s/m$, $s = 0, \dots, m - 1$ and zero otherwise, assuming that the same number of dwellings are added and dropped from the sample each month. If different panels in the same rotation group are independent, then $d(s) = 0$, but in general this will not be the case. This model is essentially the same as derived by Scott, Smith and Jones (1977). An example of an *in-for-4* rotation pattern over an eight month period is illustrated in Table 1. Different panels are denoted by different letters and the subscript indicates the number of times the panel has been included in the survey up to the time period indicated.

Table 1
Structure of *in-for-4* Rotation Pattern

Rotation Group	Time Period							
	t	$t+1$	$t+2$	$t+3$	$t+4$	$t+5$	$t+6$	$t+7$
1	a_1	a_2	a_3	a_4	b_1	b_2	b_3	b_4
2	c_4	d_1	d_2	d_3	d_4	e_1	e_2	e_3
3	f_3	f_4	g_1	g_2	g_3	g_4	h_1	h_2
4	i_2	i_3	i_4	j_1	j_2	j_3	j_4	k_1

In this case $r(2)$ is the correlation arising from say a_2 and a_4 , whereas $d(2)$ is the correlation associated with a_3 and b_1 . Binder and Hidioglou (1988) and Fuller *et al.* (1992) provided discussions of the data structure implied by some other rotation patterns.

The assumption that the variance of the sampling error series is constant implies that no major changes to the sample design or the population structure occur, at least over the effective length of the filters being considered. The assumption of stable autocorrelations, $r(s)$ and $d(s)$, for the population correlation also implies no major changes to the sample design or population. Estimates for $r(s)$ and $d(s)$ in (3) were obtained from a study by Bell (1998). The values used are from the Australian Labour Force Survey (ALFS) for the proportion of persons employed and also the proportion of persons unemployed and are shown in

Table 2. These were obtained by treating the rotation groups in the ALFS as replicates and measuring the autocorrelation at the rotation group level. A model given in Bell (1998) was used to extrapolate values beyond the given lags.

Table 2
Autocorrelations – ALFS

Proportion of employed persons								
lag	1	2	3	4	5	6	7	8
$r(s)$	0.80	0.71	0.64	0.57	0.50	0.45	0.40	0.36
$d(s)$	0.15	0.15	0.14	0.13	0.12	0.11	0.11	0.10
Proportion of unemployed persons								
lag	1	2	3	4	5	6	7	8
$r(s)$	0.62	0.52	0.44	0.37	0.31	0.26	0.22	0.19
$d(s)$	0.11	0.11	0.10	0.09	0.09	0.08	0.08	0.07

Sutcliffe and Lee (1995) studied the standard errors of seasonally adjusted and trend estimates of level and movement under a small number of different rotation patterns. They assumed a simple geometric decay model for the correlations between survey estimates with a population correlation of $\rho = 0.8$, i.e., $R(s) = \rho^s$, which decreases more rapidly than the values given in Table 2.

4. LINEAR APPROXIMATIONS FOR SEASONALLY ADJUSTED AND TREND ESTIMATES

The X11 method consists of an iterative application of moving averages resulting in a symmetric filter for the central values, and asymmetric filters for the values at the beginning and end of the series. The final seasonally adjusted and trend estimates produced by X11 can be approximated by linear filters. Several authors; for example, Young (1968), Cleveland and Tiao (1976), Wallis (1982), and Sutcliffe (1993), have produced linear approximations to the X11 procedure. The X11ARIMA procedure (Dagum 1980, 1988) is an extension of X11 and extrapolates the original series at both ends by an ARIMA model. The effect of the ARIMA extrapolation can be incorporated into the filter weights and these weights can be applied to the data alone. Dagum, Chhab and Chiu (1996) considered a Cascade method approach, where the Cascade filters are a result of the convolution of the various predetermined linear filters used within both X11 and X11ARIMA. We used this approach to realistically approximate both the X11 and X11ARIMA procedures.

Define the matrix whose rows contain the filter weights of 13 term HMAs for both symmetric and asymmetric filters as H_{13} .

The matrix of weights corresponding to a 3×3 moving average (ma) is denoted as $S_{3 \times 3}$ and that corresponding to a 3×5 ma is denoted as $S_{3 \times 5}$. These are used for estimation of seasonal factors. The matrix D is defined as a 12 term centered ma and I is an identity matrix. The notation c indicates the complement of a filter, for example $D^c = I - D$. The Seasonal Adjustment Cascade filters are written as

$$S = I - D^c S_{3 \times 5} [H_{13} (D^c S_{3 \times 3} D^c)^c]^c.$$

The trend Cascade filters used for the estimation of trend are then found by multiplying the seasonally adjusted filter by a trend filter. At the end of the series the Cascade filters for trend and seasonally adjusted estimates will differ according to whether X11 or X11ARIMA is used.

We consider the following different combinations of the internal filters of X11 and X11ARIMA:

1. Standard X11 Cascade filter: This corresponds to a 13 term HMA for estimation of trend (H_{13}), 3×3 ma for the first estimation of the seasonal factors ($S_{1 \times 3}$), 3×5 ma for estimation of seasonal factors ($S_{2 \times 5}$), and no modification for outliers.
2. Standard X11 Cascade filter with ARIMA forecasts: This corresponds to use of a H_{13} , $S_{1 \times 3}$, $S_{2 \times 5}$, and extended forecasts from an ARIMA model of the form $(1 - B)(1 - B^{12})y_t = (1 - 0.4B)(1 - 0.6B^{12})a_t$, where B is the backward shift operator and a_t is a white noise process, and no modification for outliers.
3. Short X11 Cascade filter with ARIMA forecasts: This corresponds to use of a H_9 , $S_{1 \times 3}$, $S_{2 \times 5}$, and extended forecasts from a model of the form $(1 - B)(1 - B^{12})y_t = (1 - 0.3B)(1 - 0.3B^{12})a_t$, and no modification for outliers.
4. Long X11 Cascade filter with ARIMA forecasts: This corresponds to use of a H_{23} , $S_{1 \times 3}$, $S_{2 \times 5}$, and extended forecasts from a model of the form $(1 - B)(1 - B^{12})y_t = (1 - 0.8B)(1 - 0.8B^{12})a_t$, and no modification for outliers.

Combinations 2 and 3 have been observed by Dagum (1983) to be applicable in a number of cases. The linear approximations chosen allow us to examine the effect of different rotation patterns for a range of filters used in practice, which involve HMAs of different lengths.

For each combination of filters the corresponding Cascade filter provides a vector of filter weights for the seasonally adjusted estimates and a different weight vector for the final trend estimates. These can then be substituted into equation (2) to obtain the sampling variances for a particular rotation pattern by using the appropriate values for $V(y_T | Y_T)$. When computing change estimates the data vector y_T remains unchanged and the weights that are applied change. For example, $w_{t+1} - w_t$ can be used for a one month difference. This basic approach is the same as that adopted by Wolter and Monsour (1981) who proposed estimating the variance of seasonally adjusted estimates using (2) with weights chosen that reasonably approximate the seasonal adjustment process and using a survey based estimate of $V(y_T | Y_T)$. We also consider trend filters and different realisations of X11ARIMA and rotation patterns.

The X11ARIMA models considered in this paper are representative of those commonly used in practice.

Additional complications arise from the use of ARIMA forecasts in the X11ARIMA approach. For example, we assume no misspecification of the ARIMA model. The ARIMA model is typically identified and estimated using previous survey data. The sampling error for previous time points could influence the choice of ARIMA model and X11 filters. This could be taken into consideration by modification of the variance in (2).

The initial trend and seasonally adjusted estimates for time t will be made using the time series of estimates ending at time t , that is y_t , giving the filtered value $w'_t y_t$. The value that would be obtained if there was no sampling error is $w'_t Y_t$. The sampling error considered in this paper is $w'_t y_t - w'_t Y_t$. As estimates are added to the series the filtered value for time t may change, but there will come a time point, $t + s$, after which there is no appreciable change. The final filtered value for time t based on the survey estimates can be written as $w'^*_{t+s} y_{t+s}$, for a final symmetric weight vector w'^*_t . Similarly the final value that would be obtained if there were no sampling error would be $w'^*_{t+s} Y_{t+s}$. Bell and Kramer (1999) considered the difference $w'_t y_t - w'^*_{t+s} Y_{t+s}$, which includes the forecast error. This difference can be decomposed as

$$w'_t y_t - w'^*_{t+s} Y_{t+s} = (w'_t y_t - w'_t Y_t) + (w'_t Y_t - w'^*_{t+s} Y_{t+s}).$$

We have considered how different rotation patterns affect the first term in this decomposition. The second term involves the series observed without sampling error and is unaffected by the sample design, including the rotation pattern. Bell and Kramer (1999) considered the series of US Housing Starts involving five or more units and showed that the total variance of the trend series showed large increases at the end of the series due to forecasting errors. This is due to the revisions in the initial trend estimates that are made as estimates are added to the series. Steel and McLaren (2000) considered the effect of different rotation patterns on the observed revision of the initial trend estimates, which is $w'_t y_t - w'^*_{t+s} y_{t+s}$. They noted that the relative importance of the component due to sampling error will depend on how the true series is evolving around the period being considered.

5. RESULTS

We use filters corresponding to the level and one month difference for both the seasonally adjusted and trend estimates at the very end of the series. Tables 3 to 6 summarise the effect of different rotation patterns for each Cascade filter combination. These tables give, for a selection of rotation patterns, the ratio of the sampling variance of the estimates under consideration divided by the sampling variance that would be obtained when there is complete rotation each month. The ratios obtained in the middle of the series give the same general conclusions (McLaren 1999).

Table 3

Ratio of the Sampling Variance for Chosen Rotation Patterns Divided by the Sampling Variance for an Independent Design (Combination 1)

Rotation Pattern	$\hat{S}A_t$		$\hat{S}A_{t+1} - \hat{S}A_t$		\hat{T}_t		$\hat{T}_{t+1} - \hat{T}_t$	
	emp	unemp	emp	unemp	emp	unemp	emp	unemp
complete	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1-2-1(5)	0.99	0.99	0.99	1.00	0.99	1.00	0.68	0.79
1-2-1(8)	0.98	0.99	0.97	0.99	0.98	0.99	0.64	0.77
1-1-1(6)	1.01	1.01	1.00	1.00	1.17	1.14	0.7	0.82
2-2-2(8)	1.02	1.02	0.61	0.71	1.26	1.23	0.83	0.95
2-10-2(4)	1.04	1.04	0.61	0.71	1.35	1.30	1.32	1.26
3-3-3(6)	1.07	1.06	0.48	0.61	1.52	1.44	1.29	1.25
4-8-4(8)	1.10	1.08	0.42	0.57	1.69	1.57	1.42	1.34
6-6-6(12)	1.10	1.08	0.36	0.52	1.76	1.64	1.22	1.22
in-for-6	1.10	1.08	0.36	0.52	1.76	1.64	1.22	1.22
in-for-8	1.09	1.08	0.33	0.50	1.78	1.65	1.06	1.13
no rotation	1.08	1.08	0.24	0.44	1.80	1.69	0.75	0.95

Table 4

Ratio of the Sampling Variance for Chosen Rotation Patterns Divided by the Sampling Variance for an Independent Design (Combination 2)

Rotation Pattern	$\hat{S}A_t$		$\hat{S}A_{t+1} - \hat{S}A_t$		\hat{T}_t		$\hat{T}_{t+1} - \hat{T}_t$	
	emp	unemp	emp	unemp	emp	unemp	emp	unemp
complete	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1-2-1(5)	1.01	1.01	0.99	1.00	1.06	1.05	0.69	0.80
1-2-1(8)	1.00	1.00	0.96	0.99	1.07	1.05	0.66	0.78
1-1-1(6)	1.04	1.03	1.00	1.00	1.22	1.17	0.65	0.77
2-2-2(8)	1.05	1.04	0.60	0.71	1.32	1.26	0.81	0.92
2-10-2(4)	1.02	1.03	0.60	0.71	1.26	1.23	1.19	1.17
3-3-3(6)	1.08	1.06	0.49	0.61	1.49	1.40	1.19	1.16
4-8-4(8)	1.06	1.06	0.41	0.56	1.56	1.47	1.13	1.13
6-6-6(12)	1.08	1.07	0.35	0.52	1.67	1.56	0.93	1.01
in-for-6	1.10	1.08	0.36	0.52	1.69	1.56	0.94	1.01
in-for-8	1.11	1.08	0.32	0.49	1.75	1.61	0.82	0.93
no rotation	1.14	1.11	0.24	0.43	1.89	1.73	0.59	0.78

Table 5

Ratio of the Sampling Variance for Chosen Rotation Patterns Divided by the Sampling Variance for an Independent Design (Combination 3)

Rotation Pattern	$\hat{S}A_t$		$\hat{S}A_{t+1} - \hat{S}A_t$		\hat{T}_t		$\hat{T}_{t+1} - \hat{T}_t$	
	emp	unemp	emp	unemp	emp	unemp	emp	unemp
complete	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1-2-1(5)	0.99	0.99	0.96	0.98	0.99	0.99	0.68	0.79
1-2-1(8)	0.97	0.99	0.93	0.97	0.98	0.99	0.64	0.77
1-1-1(6)	1.04	1.02	0.99	0.99	1.11	1.08	0.6	0.72
2-2-2(8)	1.07	1.06	0.60	0.71	1.23	1.19	0.89	0.95
2-10-2(4)	1.05	1.06	0.61	0.72	1.21	1.20	1.07	1.08
3-3-3(6)	1.15	1.12	0.51	0.63	1.41	1.32	1.02	1.02
4-8-4(8)	1.12	1.11	0.44	0.58	1.41	1.35	0.85	0.93
6-6-6(12)	1.14	1.13	0.37	0.53	1.47	1.39	0.69	0.82
in-for-6	1.16	1.13	0.38	0.53	1.49	1.40	0.70	0.81
in-for-8	1.17	1.14	0.34	0.51	1.52	1.42	0.61	0.76
no rotation	1.22	1.17	0.25	0.44	1.62	1.50	0.44	0.64

Table 6
Ratio of the Sampling Variance for Chosen Rotation Patterns
Divided by the Sampling Variance for an Independent Design
(Combination 4)

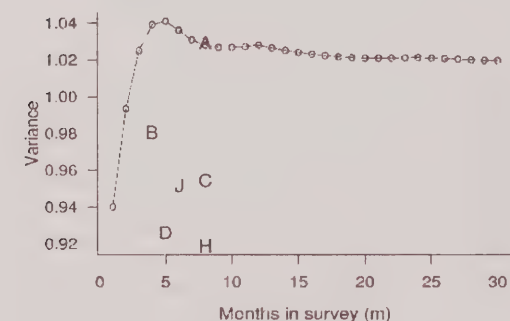
Rotation Pattern	$\hat{S}A_t$		$\hat{S}A_{t+1} - \hat{S}A_t$		\hat{T}_t		$\hat{T}_{t+1} - \hat{T}_t$	
	emp	unemp	emp	unemp	emp	unemp	emp	unemp
complete	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1-2-1(5)	1.02	1.02	0.99	1.00	1.25	1.19	0.75	0.87
1-2-1(8)	1.02	1.02	0.97	0.99	1.28	1.21	0.7	0.84
1-1-1(6)	1.06	1.04	1.00	1.00	1.49	1.39	0.92	1.01
2-2-2(8)	1.06	1.04	0.60	0.71	1.57	1.47	0.98	1.09
2-10-2(4)	1.00	1.01	0.60	0.70	1.30	1.27	1.49	1.37
3-3-3(6)	1.07	1.05	0.48	0.61	1.64	1.54	1.34	1.36
4-8-4(8)	1.05	1.04	0.41	0.56	1.73	1.63	1.92	1.69
6-6-6(12)	1.08	1.06	0.35	0.51	2.00	1.84	1.87	1.68
<i>in-for-6</i>	1.09	1.07	0.35	0.52	2.00	1.84	1.90	1.70
<i>in-for-8</i>	1.11	1.08	0.32	0.49	2.15	1.96	1.73	1.62
no rotation	1.17	1.12	0.24	0.43	2.56	2.27	1.11	1.33

5.1 X11 – Concurrent Standard Cascade Filters

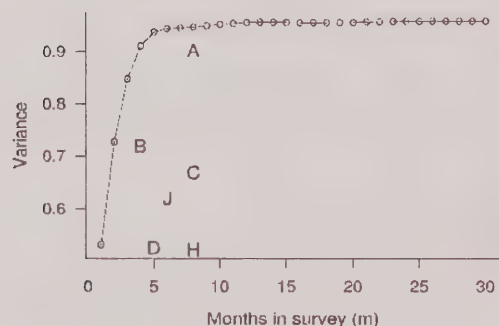
The results using the standard X11 filters (combination 1) are shown in Table 3. Figures 1(a) to 1(d) show the sampling

variance of the level and one month difference for the seasonally adjusted and trend estimates at the end of the series divided by the variance of the original estimate of level plotted against the total number of times a selected unit is included. Results for the variable employment have been plotted for selected *a-b-a(m)* patterns and the *in-for-m* rotation patterns for *m* going from 1 to 30. An *in-for-30* rotation pattern is indicative of having no rotation.

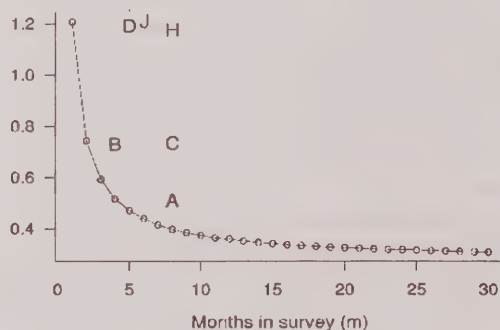
Columns 1 and 2 in Table 3 show that for the variance of the seasonally adjusted level estimates, rotation patterns with no monthly overlap perform well. Using rotation patterns with annual overlap did not help appreciably. However, for the one month change in seasonally adjusted estimates, the benefit of having high monthly overlap becomes evident (see Figure 1(b) and columns 3 and 4 of Table 3). The variances associated with the *in-for-m* rotation patterns are effectively a function of $1/m$, the proportion of the sample that does not overlap. Those rotation patterns used in Canada and Australia perform well. The best option is no rotation but, as discussed in section 2, this is not a practical option. Figures 1(a) and 1(b) show that rotation patterns that have the same degree of monthly sample overlap have similar variances for estimates of the level and one month change in the seasonally adjusted series.



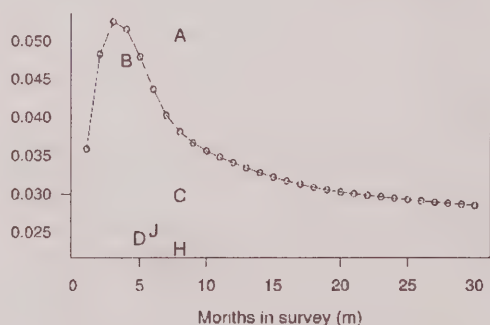
1a) Seas. adjusted: level estimates



1c) Trend: level estimates



1b) Seas. adjusted: one month change



1d) Trend: one month change

Figure 1. Ratio of the sampling variance to the variance of the original series for chosen rotation patterns for combination 1 (X11) for the variable employment where A = 4-8-4(8), B = 2-10-2(4), C = 2-2-2(8), D = 1-2-1(5), H = 1-2-1(8), J = 1-1-1(6).

For the level of trend estimates the variance increases as the amount of monthly sample overlap increases (see Figure 1(c) and columns 5 and 6 of Table 3). For the *in-for-m* rotation patterns there is a rapid increase in variance as m goes from 1 to 5. The rotation patterns of 1-2-1(5) and 1-2-1(8) perform as well as having an independent sample each month and considerably better than rotation patterns that involve monthly overlap. This is primarily due to the fact that for a moving average, it is better to average over independent observations than positively correlated ones. The larger variance of the 1-1-1(6) pattern compared with that of 1-2-1(5) and 1-2-1(8) suggest that, for those patterns with no monthly overlap, the interval between the re-inclusion of units in the sample has some effect.

Figure 1(d) and columns 7 and 8 of Table 3, show that for one month changes in trend estimates the variance increases very rapidly as m increases from 1 to 3 and decreases rapidly as m increases from 4. The *in-for-3* rotation pattern seems to be the worst among those considered, and the currently used rotation patterns can be significantly improved upon. For example, using a 1-2-1(8) instead of a 4-8-4(8) rotation pattern would reduce the variance in the one month change in trend estimates for employment by 55 percent and 43 percent for unemployment. While the degree of monthly overlap is still a key factor, the pattern of inclusion also plays a role, for example the 2-2-2(8) pattern has lower variance than the *in-for-2* or 2-10-2(4) patterns. Moreover, for one month changes in the trend estimates the best performing rotation patterns are 1-2-1(5) and 1-2-1(8) which perform considerably better than using complete rotation each month. This result arises because one month changes in trend estimates effectively look at differences in the seasonally adjusted series a few months apart and the 1-2-1(m) rotation patterns lead to positive correlations between estimates 3 months apart. Similar results were obtained in a study by McLaren and Steel (1997) using Sutcliffe's (1993) approximation to X11.

The results show that for the estimation of the current level of trend and the latest movement in trend, the 1-2-1(m) rotation patterns give considerably lower sampling variances than the rotation patterns currently in use.

5.2 X11ARIMA – Concurrent Cascade Filters with Extrapolations

Results for the filter combinations 2, 3 and 4 are given in Tables 4, 5 and 6 respectively. Figures 2(a) to 2(d) present results for combination 4 for employment.

Columns 1 and 2 of Tables 4, 5 and 6 show that rotation patterns with low monthly overlap perform almost as well as complete rotation for seasonally adjusted level estimates. Rotation patterns with high monthly overlap have higher variances, particularly for combination 3 which corresponds to the use of the 9 term HMA.

There is minimal difference between the ratios of the four different combinations for the one month change in the seasonally adjusted estimates (columns 3 and 4 in all

tables). Rotation patterns with high monthly sample overlap still perform better than those with low or no monthly overlap regardless of the X11/X11ARIMA combination used.

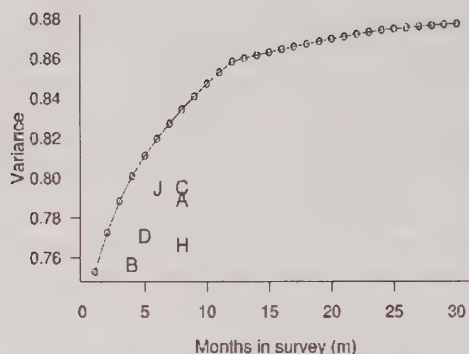
For the level of trend estimates, rotation patterns with a higher degree of sample overlap again have a greater variance ratio. The 1-2-1(5) and 1-2-1(8) rotation patterns still out-perform the other rotation patterns for each combination of filters, although they do not perform as well as an independent sample for combinations 2 and 4.

For one month changes in the trend estimates the better performing rotation patterns are again 1-2-1(5) and 1-2-1(8) which perform better than the independent sample for all four combinations of filters. For combination 3, rotation patterns with high monthly overlap perform equally as well as the 1-2-1(m) rotation patterns. For combinations 2 and 3 the 1-1-1(6) pattern is slightly better than the 1-2-1(m) patterns. Substantial improvements over the currently used rotation patterns can be achieved by using 1-2-1(m) rotation patterns. For example, for the employment variable, changing from an 4-8-4(8) to a 1-2-1(8) would produce gains of 42, 25 and 64 percent using combinations 2, 3 and 4, respectively.

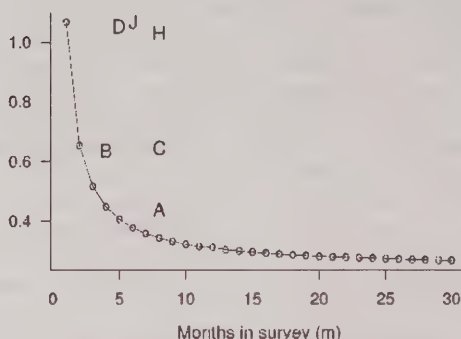
These results are based on the ALFS correlation estimates which, being based on survey estimates, will be subject to sampling error. The trend filters considered are not derived using these estimates. The same general conclusions concerning the impact of different rotation patterns are obtained for the two correlation models which use reasonably different correlations. We believe that the conclusions will apply for the range of correlation models contained between these two models. Similar conclusions are also obtained by McLaren and Steel (1997) using a correlation model derived by Steel (1996) for UK employment and unemployment.

6. DISCUSSION

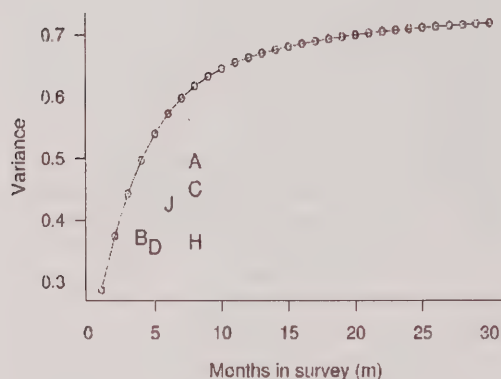
The rotation patterns currently used, such as *in-for-8*, *in-for-6* and 4-8-4(8), are sensible if the one month change in seasonally adjusted estimates are the key statistics to be analysed. We believe that examination of the one month change in seasonally adjusted estimates is often not a reliable way of assessing current trends. It is necessary to look at the pattern of change over recent months. This can be done using filters to obtain an estimate of the trend. The results here suggest if the main use of the survey is to provide an assessment of trend then quite different rotation patterns should be used. Specifically, the 1-2-1(m) rotation patterns performed well for reducing the variance of the level of trend estimates and the difference between two consecutive trend estimates for a range of different filter combinations. The 1-2-1(m) rotation patterns also performed well for the sampling variance of the seasonally adjusted level estimates. Hence, in designing the rotation pattern for a repeated survey, the relative importance of



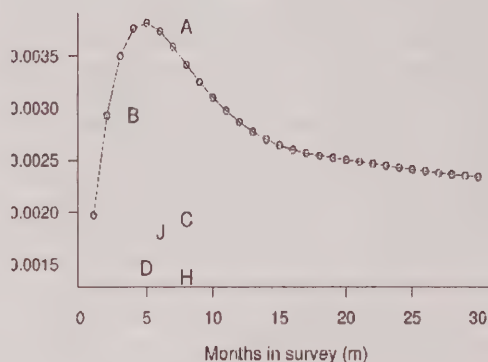
2a) Seas. adjusted: level estimates



2b) Seas. adjusted: one month change



2c) Trend: level estimates



2d) Trend: one month change

Figure 2. Ratio of the sampling variance to the variance of the original series for chosen rotation patterns for Combination 4 (X11ARIMA) for the variable employment where A = 4-8-4, B = 2-10-2(4), C = 2-2-2(8), D = 1,2-1(5), H = 1-2-1(8), J = 1-1-1(6).

seasonally adjusted and trend estimates needs to be carefully considered. Examining Figures 1 and 2 shows that the rotation pattern 2-2-2(8), is a reasonable compromise if the level and one months change in seasonally adjusted and trend estimates are both considered important. Bell (1999) also considered the effect of four different rotation patterns on the sampling variance of the level and one month change in the original, unadjusted, estimates and also trend estimates obtained using X11 and a 13 point HMA. He also identifies the 2-2-2(8) rotation pattern as a compromise design.

Even if analysts do not formally use trend estimates, the assessment of trend will involve looking at changes in seasonally adjusted estimates a few months apart. McLaren (1999) gives results which show that the 1-2-1(*m*) rotation patterns will be suitable if the assessment of trends involve looking at changes in seasonally adjusted estimates over 3 or 6 months. The results also suggest that such rotation patterns perform well for estimates of the change in trend estimates over the most recent 3 and 6 months.

The evaluation criterion used in this paper is the sampling variance of the trend and seasonally adjusted estimates, which is the factor affected by the sample design. Steel and McLaren (2000) considered assessing different rotation patterns in terms of the degree of revisions of these estimates at the end points and reached similar conclusions regarding the rotation patterns.

ACKNOWLEDGEMENT

This research was supported by the Australian Research Council and the Australian Bureau of Statistics (ABS). The views expressed in this paper may not necessarily reflect the views of either organisation. We would like to thank the associate editor and the referees for their comments and Geoff Lee, Andrew Sutcliffe and Phillip Bell from the ABS, and Norma Chhab from Statistics Canada.

REFERENCES

- AUSTRALIAN BUREAU OF STATISTICS (1987). *A Guide to Smoothing Time Series – Estimates of "Trend"*. Australian Bureau of Statistics, catalogue no. 1316.0, Canberra.
- AUSTRALIAN BUREAU OF STATISTICS (1992). *Information Paper: Labour Force Survey Sample Design*. Australian Bureau of Statistics, catalogue no. 6269.0, Canberra.
- AUSTRALIAN BUREAU OF STATISTICS (1993). *A Guide to Interpreting Time Series – Monitoring "Trends", An Overview*. Australian Bureau of Statistics, catalogue no. 1348.0, Canberra.
- BELL, P.A. (1998). *Using State Space Models and Composite Estimation to Measure the Effects of Telephone Interviewing on Labour Force Estimates*. Working Papers in Econometrics and Applied Statistics, No. 98/2, Australian Bureau of Statistics, catalogue no. 1351.0, Canberra.
- BELL, P.A. (1999). *The Impact of Sample Rotation Patterns and Composite Estimation on Survey Outcomes*. Working Papers in Econometrics and Applied Statistics, No. 99/1, Australian Bureau of Statistics, catalogue no. 1352.0, Canberra.
- BELL, W.R., and HILLMER, S. (1990). Time series methods for survey estimation. *Survey Methodology*, 16, 195-215.
- BELL, W.R., and KRAMER, M. (1999). Towards variances for X-11 seasonal adjustment. *Survey Methodology*, 25, 13-29.
- BELL, W.R., and WILCOX, D.W. (1993). The effect of sampling error on the time series behavior of consumption data. *Journal of Econometrics*, 555, 235-265.
- BINDER, D.A., and HIDIROGLOU, M.A. (1988). Sampling in time. *Handbook of Statistics*, (P.R. Krishnaiah and C.R. Rao, Eds.), Amsterdam: Elsevier Science Publishers, B.V., 6, 187-211.
- BURRIDGE, P., and WALLIS, K.F. (1985). Calculation of seasonally adjusted series. *Journal of the American Statistical Association*, 80, 541-552.
- CLEVELAND, W.P., and TIAO, G.C. (1976). Decomposition of seasonal time series: a model for the X-11 program. *Journal of the American Statistical Association*, 71, 581-587.
- DAGUM, E.B. (1980). *The X-11ARIMA Seasonal Adjustment Method*. Catalogue no. 12-564E, Statistics Canada, Ottawa.
- DAGUM, E.B. (1983). Spectral properties of the concurrent and forecasting seasonal linear filters of the X-11ARIMA method. *The Canadian Journal of Statistics*, 11, 73-90.
- DAGUM, E.B. (1988). *The X-11ARIMA/88 Seasonal Adjustment Methods – Foundations and User's Manual*. Statistics Canada, Ottawa.
- DAGUM, E.B., CHHAB, N., and CHIU, K. (1996). Derivation and properties of the X-11ARIMA and Census X-11 linear filters. *Journal of Official Statistics*, 12, 329-347.
- DEMPTSTER, P.A., and HWANG, J.-S. (1993). Component models and Bayesian technology for estimation of state employment and unemployment rates. *Proceedings of the Bureau of the Census Annual Research Conference*, 571-581.
- FINDLEY, D.F., MONSELL, B.C., OTTO, M.C., BELL, W.R., and PUGH, M.G. (1998). New capabilities and methods of the X-12 ARIMA seasonal adjustment program. *Journal of Business and Economic Statistics*, 16, 127-177.
- FULLER, W.A., ADAM, A., and YANSANEH, I.S. (1992). Estimates for longitudinal surveys with applications to the U.S. Current Population Survey. *Proceedings: Symposium 92, Design and Analysis of Longitudinal Surveys*, Statistics Canada, 301-324.
- GRAY, A., and THOMSON, P. (1996). Design of moving-average trend filters using fidelity, smoothness and minimum revisions criteria. *Time Series Analysis in Memory of E.J. Hannan*. (Ed. P. Robinson and M. Rosenblatt), 205-219. Springer lecture notes in statistics, 115.
- HENDERSON, R. (1916). Note on graduation by adjusted averages. *Transactions of the Actuarial Society of America*, 17, 43-48.
- HAUSMAN, J.A., and WATSON, M.W. (1985). Error in variables and seasonal adjustment procedures. *Journal of the American Statistical Association*, 80, 531-540.
- KALTON, G., and CITRO, C.F. (1993). Panels surveys: adding the fourth dimension. *Survey Methodology*, 19, 205-215.
- KENNY, P.B., and DURBIN, J. (1982). Local trend estimation and seasonal adjustment of economic and social time series. *Journal of the Royal Statistical Society A*, 145, 1-41.
- KISH, L. (1998). Space/time variations and rolling samples. *Journal of Official Statistics*, 14, 31-46.
- KNOWLES, J. (1997). Trend Estimation Practices of National Statistical Institutes. Office for National Statistics, Methods and Quality Division, UK, MQ 044.
- KNOWLES, J., and KENNY (1997). An Investigation of Trend Estimation Methods. Office for National Statistics, Methods and Quality Division, UK, MQ 044.
- LEE, H. (1990). Estimation of panel correlation for the Canadian Labour Force Survey. *Survey Methodology*, 16, 283-292.
- LINACRE, S., and ZARB, J. (1991). Picking turning points in the economy. *Australian Economic Indicators*, Australian Bureau of Statistics, catalogue no. 1350.0.
- MCLAREN, C.H. (1999). Designing Rotation Patterns and Filters for Trend Estimation in Repeated Surveys. Unpublished PhD Thesis, School of Mathematics and Applied Statistics, University of Wollongong.
- MCLAREN, C.H., and STEEL, D.G. (1997). The effect of different rotation patterns on the sampling variance of seasonal and trend filters. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1997, 790-795.
- MAIAZAKI, E.S., and DOREA, C.C.Y. (1993). Estimation of the parameters of a time series subject to the error of rotation sampling. *Communications in Statistics, A*, 22, 805-825.
- PFEFFERMANN, D. (1994). A general method for estimating the variances of X-11 seasonally adjusted estimators. *Journal of Time Series Analysis*, 15, 85-116.
- RAO, J.N.K., and GRAHAM, J.E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 69, 492-509.
- SCOTT, A.J., SMITH T.M.F., and JONES, R. G. (1977). The application of time series methods to the analysis of repeated surveys. *International Statistical Review*, 45, 3-73.

- SHISKIN, J., YOUNG, A.H., and MUSGRAVE, J.C. (1967). *X-11 Variant of the Census Method II Seasonal Adjustment Program*. Technical Paper 15, Bureau of the Census, U.S. Department of Commerce, Washington, D.C.
- SINGH, M.P., DREW, J.D., GAMBINO, J., and MAYDA, F. (1990). *Methodology of the Canadian Labour Force Survey*. Catalogue no. 71-526, Statistics Canada.
- SMITH, T.M.F. (1997). Discussion of paper by Steel. *Journal of the Royal Statistical Society A*, 160, 33-34.
- STEEL, D.G. (1996). Options for Producing Monthly Estimates of Unemployment According to the ILO Definition. Central Statistical Office, U.K.
- STEEL, D.G. (1997). Producing monthly estimates of unemployment and employment according to the international labour office definition. *Journal of the Royal Statistical Society A*, 160, 5-46.
- STEEL, D.G., and MCLAREN, C.H. (2000). The effect of different rotation patterns on the revisions of trend estimates. *Journal of Official Statistics*, 16, 61-76.
- STEEL, D.G., and DEMEL, R. (1988). The Contribution of Sampling Error to the Variability of Statistical Series. Paper Presented at the National Mathematical Sciences Congress, Canberra.
- SUTCLIFFE, A. (1993). *X-11 Time Series Decomposition and Sampling Errors*. Working Papers in Econometrics and Applied Statistics, No 93/2. Australian Bureau of Statistics, catalogue no 1351.
- SUTCLIFFE, A., and LEE, G. (1995). Seasonal Analysis and Sample Design. Paper presented at the Conference of Survey Measurement and Process Quality. Bristol 1995.
- TILLER, R.B. (1992). Time series modeling of sample survey data from the U.S. Current Population Survey. *Journal of Official Statistics*, 8, 149-166.
- WALLIS, K.F. (1982). Seasonal adjustment and revision of current data: linear filters for the X-11-method. *Journal of the Royal Statistical Society A*, 145, 74-85.
- WOLTER, K.M., and MONSOUR, N.J. (1981). On the problem of variance estimation for a deseasonalized series. *Current Topics in Survey Sampling*, (D. Krewski, R. Platek and J.N.K. Rao, Eds.). New York: Academic Press, 367-407.
- YOUNG, A.H. (1968). Linear Approximations to the Census and BLS Seasonal Adjustment Methods. *Journal of the American Statistical Association*, 63, 445-471.

Hierarchical Bayes Estimation of Small Area Means Using Multi-Level Models

YONG YOU and J.N.K. RAO¹

ABSTRACT

Standard multi-level models with random regression parameters are considered for small area estimation. We also extend the models by allowing unequal error variances or by assuming random effect models for both regression parameters and error variances. We present these models in a hierarchical Bayes framework and estimate a small area mean by its posterior mean. Posterior variance of the small area mean is used as a measure of precision of the estimate. It automatically takes into account the extra uncertainty associated with the hyperparameters in the multi-level model. Gibbs sampling is used to compute the posterior means and posterior variances of small area means. Rao-Blackwellized estimators that reduce the Monte Carlo errors are obtained. Bayesian model selection and sensitivity analysis are also studied. The procedure is illustrated using data on household income in some counties (small areas) of Brazil.

KEY WORDS: Gibbs sampling; Hierarchical Bayes; Multi-level model; Sampling error variance; Small area.

1. INTRODUCTION

Small area estimation has received a lot of attention in recent years due to growing demand for reliable small area estimators. Traditional area-specific direct estimators do not provide adequate precision because sample sizes in small areas are seldom large enough. This makes it necessary to employ indirect estimators that borrow strength from related areas; in particular, model-based indirect estimators. Battese, Harter and Fuller (1988) proposed and applied a nested error regression model to provide model-based small area estimates. The model takes the form

$$y_{ij} = x_{ij}^T \beta + v_{0i} + e_{ij}, j = 1, \dots, n_i; i = 1, \dots, m, \quad (1)$$

where y_{ij} are the observations associated with the sampled units in the i -th small area, $i = 1, \dots, m$, x_{ij} is the $p \times 1$ vector of unit-level explanatory variables, β is a set of p fixed regression parameters, v_{0i} are independent area effects with $E(v_{0i}) = 0$ and $V(v_{0i}) = \sigma_v^2$. The e_{ij} 's are assumed to be independent random error variables with $E(e_{ij}) = 0$ and $V(e_{ij}) = \sigma_e^2$. v_{0i} and e_{ij} are also assumed to be independent. For the whole population, model (1) applies with n_i replaced by N_i , the small area population size. The model (1) may be expressed in matrix notation as follows

$$Y_i = X_i \beta + v_{0i} \mathbf{1}_{n_i} + e_i, i = 1, \dots, m,$$

where $Y_i = (y_{i1}, \dots, y_{i, n_i})^T$, $X_i = (x_{i1}, \dots, x_{i, n_i})^T$ is a $n_i \times p$ matrix, $\mathbf{1}_{n_i} = (1, \dots, 1)^T$ is the unit vector of length n_i , and $e_i = (e_{i1}, \dots, e_{i, n_i})^T$.

Holt and Moura (1993) extended the above framework to a multi-level model by introducing random regression coefficients and then relating them to area-level explanatory

variables to explain some of the between small area variation. The model can be stated as follows:

$$Y_i = X_i \beta_i + e_i, \beta_i = Z_i \gamma + v_i \quad (2)$$

where Z_i is the $p \times q$ design matrix of area-level variables, γ is a $q \times 1$ vector of fixed coefficients, and $v_i = (v_{i1}, \dots, v_{ip})^T$ is a $p \times 1$ vector of random effects for the i -th area. The v_i 's are independent across areas and have a joint distribution within each area with $E(v_i) = 0$ and $V(v_i) = \Phi$, where the variance covariance matrix Φ is unknown. Note that model (2) effectively integrates the use of unit-level and area-level covariates into a single model. Holt and Moura (1993) and Moura and Holt (1999) extended Prasad and Rao's (1990) framework to the above multi-level model to get the best linear unbiased predictor (BLUP) of the small area mean $\mu_i = \bar{X}_i^T \beta_i$ assuming that N_i is large, where \bar{X}_i is the $p \times 1$ vector of known population means of the auxiliary variables for the i -th small area. They also obtained the empirical BLUP (EBLUP) and a second order approximation to the mean squared error (MSE) of EBLUP for the multi-level model. Using household income data in some counties (small areas) of Brazil, they demonstrated gain in efficiency of the EBLUP estimators over the EBLUP estimators obtained from nested error regression models. Ghosh and Rao (1994) and Rao (1999) provide a detailed overview of model-based methods for small area estimation.

In this paper, we study the multi-level model (2) in a hierarchical Bayes framework and extend the model to more general multi-level models which allow fixed unequal error variances or random error variances. The small area mean μ_i is estimated by its posterior mean and its precision is measured by its posterior variance. Posterior variance automatically takes into account the extra uncertainty

¹ Yong You, Household Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6; J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada, K1S 5B6.

associated with the hyperparameters in the multi-level model. We use the Gibbs sampling method to compute the hierarchical Bayes estimates and the associated posterior variances. Section 2 presents the hierarchical Bayes multi-level models with different assumptions on error variances and related Gibbs sampling inference. Section 3 illustrates our methodology and studies model selection and sensitivity analysis by employing data on household incomes in some counties (small areas) of Brazil. And finally in section 4, we give some comments and concluding remarks.

2. MULTI-LEVEL MODELS AND GIBBS SAMPLING INFERENCE

2.1 Equal Error Variances

We consider a hierarchical Bayes representation of the multi-level model (2) as follows:

Model 1:

(i) Conditional on β_j and σ_e^2 , y_{ij} 's are independent with

$$y_{ij} | \beta_j, \sigma_e^2 \sim N(x_{ij}^T \beta_j, \sigma_e^2), \quad (i = 1, \dots, m; j = 1, \dots, n_j); \quad (3)$$

(ii) Conditional on γ and Φ , β_i 's are independent with

$$\beta_i | \gamma, \Phi \sim N_p(Z_i \gamma, \Phi), \quad (i = 1, \dots, m). \quad (4)$$

To complete our Bayesian model specification, we adopt the prior distributions for parameters as follows:

(iii) Marginal prior distributions: $\gamma \sim N_q(0, D)$, $\tau_e \sim G(a, b)$ and $\Omega \sim W_p(\alpha, R)$, where $\tau_e = \sigma_e^2$, $\Omega = \Phi^{-1}$, and D, a, b, α and R are known.

In step (iii) of Model 1, $G(a, b)$ denotes a gamma distribution with density given by $f(x) = b^a / \Gamma(a) x^{a-1} e^{-xb}$, $a > 0, b > 0, x \geq 0$, and $W_p(\alpha, R)$ is a Wishart distribution with density function

$$f(X) = \frac{|R|^{\frac{\alpha}{2}}}{2^{ap/2} \Gamma_p\left(\frac{\alpha}{2}\right)} |X|^{-\frac{\alpha-p-1}{2}} \exp\left\{-\frac{1}{2} \text{tr}(RX)\right\},$$

where $X > 0, R > 0$ and $\Gamma_p(\alpha)$ is multivariate gamma function defined as

$$\Gamma_p(\alpha) = \pi^{\frac{p(p-1)}{4}} \prod_{j=1}^p \Gamma\left(\alpha + \frac{1}{2}(1-j)\right).$$

Remark 1.1: The prior distributions in step (iii) are conjugate with the sampling and population distributions given by (3) and (4) in the sense that they lead to full conditional distributions for γ , τ_e and Ω that are again normal, gamma and Wishart distribution, respectively. The Wishart distribution is the multivariate version of gamma distribution for

the inverse variance covariance matrix of random effects. The importance of conjugacy may be exploited as follows: (1) In the Gibbs sampling step, without conjugacy the full conditional distribution for any parameter will be known up to normalizing constants. In this case, more sophisticated random generation will be required. (2) Closed-form full conditional distributions may be employed to find the Rao-Blackwellized estimators of the posterior means and posterior variances, and thus to improve posterior estimation. In general, for Bayesian inference, choosing priors is not a simple job because any proper prior on the model parameters is a plausible candidate. This is a limitation of Bayesian methods.

Remark 1.2: It is important to note that we have used proper priors on all the unknown parameters to ensure that all the posterior distributions are proper (Hobert and Casella 1996). Hence we do not face the problem of some posteriors being improper. Values for the parameters of the priors (i.e., hyperparameters) are chosen to reflect a fairly vague knowledge of the prior distributions. Details will be given in section 3 on data analysis.

Remark 1.3: In Model 1, we assume equal error variance σ_e^2 for all small areas. In practice, however, variances of sampling error could be different for different small areas. A more general model should allow possibly different error variances. In sections 2.2 and 2.3, we will introduce unequal error variance and random error variance models.

We are interested in finding the posterior distributions of β_i 's given the data $Y = (\{y_{ij}\}, i = 1, \dots, m; j = 1, \dots, n_j)$, and in particular in finding the posterior estimates of small area means $\mu_i = \bar{X}_i^T \beta_i$, which depend on the estimates of β_i . Direct evaluation of the joint posterior distribution involves high-dimensional numerical integration, and is not computationally feasible. Therefore, we use the Gibbs sampling method (Gelfand and Smith 1990) to generate samples from the joint posterior distributions. To implement the Gibbs sampling under Model 1, we need the full conditional distributions given by:

- (i) For $i = 1, \dots, m$,

$$[\beta_i | Y, \gamma, \Omega, \tau] \propto N_p((\tau_i X_i^T X_i + \Omega)^{-1} (\tau_i X_i^T Y_i + \Omega Z_i \gamma), (\tau_i X_i^T X_i + \Omega)^{-1})$$
- (ii) $[\gamma | Y, \beta, \Omega, \tau_e] \sim N_q\left(\left(\sum_{i=1}^m Z_i^T \Omega Z_i + D\right)^{-1} \left(\sum_{i=1}^m Z_i^T \Omega \beta_i\right), \left(\sum_{i=1}^m Z_i^T \Omega Z_i + D\right)^{-1}\right)$
- (iii) $[\Omega | Y, \beta, \gamma, \tau] \sim W_p\left(\alpha + m, R + \frac{1}{2} \sum_{i=1}^m (\beta_i - Z_i \gamma)(\beta_i - Z_i \gamma)^T\right)$
- (iv) $[\tau_e | Y, \beta, \gamma, \Omega] \sim G\left(a + \frac{1}{2} \sum_{i=1}^m n_i, b + \frac{1}{2} \left(\sum_{i=1}^m (Y_i - X_i \beta_i)^T (Y_i - X_i \beta_i)\right)\right)$

Since all the full conditional distributions have closed-form, it is easy to generate samples. Gibbs sampling method is as follows: (a) Using starting values $\gamma^{(0)}$, $\Omega^{(0)}$ and $\tau_e^{(0)}$, draw $\beta_i^{(1)}$, $i = 1, \dots, m$, from $[\beta_i | Y, \gamma, \Omega, \tau_e]$; (b) Draw $\gamma^{(1)}$ from $[\gamma | Y, \beta, \Omega, \tau_e]$ using $\beta_i^{(1)}$, $i = 1, \dots, m$, $\Omega^{(0)}$ and $\tau_e^{(0)}$; (c) Draw $\Omega^{(1)}$ from $[\Omega | Y, \beta, \gamma, \tau_e]$ using $\beta_i^{(1)}$, $i = 1, \dots, m$, $\gamma^{(1)}$ and $\tau_e^{(0)}$; (d) Draw $\tau_e^{(1)}$ from $[\tau_e | Y, \beta, \gamma, \Omega]$ using $\beta_i^{(1)}$, $i = 1, \dots, m$, $\gamma^{(1)}$ and $\Omega^{(1)}$. Steps (a)-(d) complete one sampling cycle. Perform a large number of cycles, say t , called "burn-in" period, until convergence and then treat $\{\beta_i^{(t+k)}, i = 1, \dots, m; \gamma^{(t+k)}; \Omega^{(t+k)}, \tau_e^{(t+k)}; k = 1, \dots, G\}$ as G samples from the joint posterior of β_i , $i = 1, \dots, m$, γ , Ω and τ_e .

Suppose a sample of size G is obtained as $\{\beta_i^{(k)}, i = 1, \dots, m; \gamma^{(k)}; \Omega^{(k)}; \tau_e^{(k)}; k = 1, \dots, G\}$. To obtain an estimator of the posterior mean of β_i , one can use the sample mean of the $\{\beta_i^{(k)}\}$. Since β_i has a closed form full conditional distribution, we can use the sample mean of the conditional expectations $\{E[\beta_i | Y, \gamma^{(k)}, \Omega^{(k)}, \tau_e^{(k)}]\}$ to improve our estimation, since $E(\beta_i | Y) = E(E(\beta_i | Y, \gamma, \Omega, \tau_e))$, and $\text{Var}(\beta_i | Y) \geq \text{Var}(E(\beta_i | Y, \gamma, \Omega, \tau_e))$. This modification is based on the well-known Rao-Blackwell theorem and the corresponding estimator is the so-called Rao-Blackwellized estimator (Gelfand and Smith 1990, 1991). Thus we have the following two alternative estimators for β_i :

$$\hat{\beta}_i^{(E)} = \frac{1}{G} \sum_{k=1}^G \beta_i^{(k)} \quad (5)$$

and

$$\begin{aligned} \hat{\beta}_i^{(RB)} &= \frac{1}{G} \sum_{k=1}^G E(\beta_i | Y, \gamma^{(k)}, \Omega^{(k)}, \tau_e^{(k)}) \\ &= \frac{1}{G} \sum_{k=1}^G (\tau_e^{(k)} X_i^T X_i + \Omega^{(k)})^{-1} \\ &\quad (\tau_e^{(k)} X_i^T Y_i + \Omega^{(k)} Z_i \gamma^{(k)}), \end{aligned} \quad (6)$$

where $\hat{\beta}_i^{(E)}$ is the empirical estimator and $\hat{\beta}_i^{(RB)}$ is the Rao-Blackwellized estimator. Both $\hat{\beta}_i^{(E)}$ and $\hat{\beta}_i^{(RB)}$ are unbiased for the posterior mean. However, $\hat{\beta}_i^{(RB)}$ is better than $\hat{\beta}_i^{(E)}$ in terms of simulation standard error (Gelfand and Smith 1991).

The corresponding estimators for the small area mean μ_i are given as

$$\hat{\mu}_i^{(E)} = \bar{X}_i^T \hat{\beta}_i^{(E)} = \frac{1}{G} \sum_{k=1}^G \bar{X}_i^T \beta_i^{(k)} \quad (7)$$

and

$$\begin{aligned} \hat{\mu}_i^{(RB)} &= \bar{X}_i^T \hat{\beta}_i^{(RB)} = \frac{1}{G} \sum_{k=1}^G \bar{X}_i^T (\tau_e^{(k)} X_i^T X_i + \Omega^{(k)})^{-1} \\ &\quad (\tau_e^{(k)} X_i^T Y_i + \Omega^{(k)} Z_i \gamma^{(k)}). \end{aligned} \quad (8)$$

We anticipate that both $\hat{\mu}_i^{(E)}$ and $\hat{\mu}_i^{(RB)}$ will give almost the same point estimates. However, it will be of interest to compute and compare the simulation standard errors of these two estimators to evaluate the effects of Rao-Blackwellization; see section 3.

To obtain the posterior variance of μ_i , we first find the posterior variance of β_i , since $V(\mu_i | Y) = \bar{X}_i^T V(\beta_i | Y) \bar{X}_i$. Note that

$$\begin{aligned} V(\beta_i | Y) &= E(V(\beta_i | Y, \gamma, \Omega, \tau_e)) + V(E(\beta_i | Y, \gamma, \Omega, \tau_e)) \\ &= E(V(\beta_i | Y, \gamma, \Omega, \tau_e)) + E(E(\beta_i | Y, \gamma, \Omega, \tau_e)^2) \\ &\quad - [E(E(\beta_i | Y, \gamma, \Omega, \tau_e))]^2. \end{aligned} \quad (9)$$

Using (9), the Rao-Blackwellized estimator of the posterior variance of β_i , denoted by $\hat{V}(\beta_i)$, can be obtained using the Gibbs samples $\{\beta_i^{(k)}, i = 1, \dots, m; \gamma^{(k)}; \Omega^{(k)}; \tau_e^{(k)}; k = 1, \dots, G\}$; see Appendix A1. The posterior variance of small area mean μ_i is then estimated by

$$\hat{V}(\mu_i) = \bar{X}_i^T \hat{V}(\beta_i) \bar{X}_i. \quad (10)$$

The same estimation procedure can be applied to the sampling error variance σ_e^2 . Since conditionally σ_e^2 has an inverse gamma distribution, the Rao-Blackwellized estimator of the posterior mean of σ_e^2 is obtained as

$$\begin{aligned} \hat{\sigma}_e^{2(RB)} &= \frac{1}{G} \sum_{k=1}^G \left[b + \frac{1}{2} \sum_{i=1}^m (Y_i - X_i \beta_i^{(k)})^T (Y_i - X_i \beta_i^{(k)}) \right] \\ &\quad \left(a + \frac{1}{2} \sum_{i=1}^m n_i - 1 \right)^{-1}. \end{aligned} \quad (11)$$

Since we are mainly interested in estimating the small area means, calculation of the sampling variance is only for the purpose of model selection. Details on model selection will be given in section 3.2.

2.2 Unequal Error Variances

In practice, it is more realistic to allow unequal error variances for the sampling errors. Let σ_i^2 be the true sampling error variance for the i -th small area. A straightforward extension of Model 1 leads to the following hierarchical Bayes multi-level unequal error variance model:

Model 2:

(i) Conditional on β_i and σ_i^2 , y_{ij} 's are independent with

$$y_{ij} | \beta_i, \sigma_i^2 \sim N(x_{ij}^T \beta_i, \sigma_i^2), \quad (i = 1, \dots, m; j = 1, \dots, n_i); \quad (12)$$

(ii) Conditional on γ and Φ , β_i 's are independent with $\beta_i | \gamma, \Phi \sim N_p(Z_i \gamma, \Phi)$, ($i = 1, \dots, m$); (13)

- (iii) Marginal prior distributions: $\gamma \sim N_q(0, D)$, $\tau_i \stackrel{\text{iid}}{\sim} G(a_i, b_i)$, and $\Omega \sim W_p(\alpha, R)$, where $\tau_i = \sigma_i^{-2}$, $\Omega = \Phi^{-1}$, and D, a_i, b_i, α and R are known.

Remark 2.1: Model 2 reduces to Model 1 when $\sigma_i^2 = \sigma_e^2$ for all i . From a hierarchical Bayes perspective, extension from the equal error variance model to the unequal error variance model is straightforward. Also there is no difficulty in the Gibbs sampling implementation.

Remark 2.2: τ_i 's are assumed to be independent and have prior distributions $G(a_i, b_i)$, where a_i and b_i are known hyperparameters and usually chosen to be very small to reflect a vague knowledge about τ_i 's.

The full conditional distributions for Gibbs sampling under Model 2 are given by:

- (i) For $i = 1, \dots, m$,

$$[\beta_i | Y, \gamma, \Omega, \tau] \stackrel{\text{iid}}{\sim} N_p((\tau_i X_i^T X_i + \Omega)^{-1} (\tau_i X_i^T Y_i + \Omega Z_i^T \gamma), (\tau_i X_i^T X_i + \Omega)^{-1})$$
- (ii)
$$[\gamma | Y, \beta, \Omega, \tau] \sim N_q \left(\left(\sum_{i=1}^m Z_i^T \Omega Z_i + D^{-1} \right)^{-1} \left(\sum_{i=1}^m Z_i^T \Omega \beta_i \right), \left(\sum_{i=1}^m Z_i^T \Omega Z_i + D^{-1} \right)^{-1} \right)$$
- (iii)
$$[\Omega | Y, \beta, \gamma, \tau] \sim W_p \left(\alpha + m, R + \frac{1}{2} \sum_{i=1}^m (\beta_i - Z_i \gamma)(\beta_i - Z_i \gamma)^T \right)$$
- (iv) For $i = 1, \dots, m$,

$$[\tau_i | Y, \beta, \gamma, \Omega] \sim G(a_i + \frac{1}{2} n_i, b_i + \frac{1}{2} \times (Y_i - X_i \beta_i)^T (Y_i - X_i \beta_i)).$$

For Model 2, the empirical estimators of the posterior means of β_i and μ_i have the same form as (5) and (7). The Rao-Blackwellized estimators $\hat{\beta}_i^{(\text{RB})}$ and $\hat{\mu}_i^{(\text{RB})}$ are the estimators given by (6) and (8) with $\tau_e^{(k)}$ replaced by $\tau_i^{(k)}$. Estimator of posterior variance is $\hat{V}(\mu_i)$ given by (10) with $\tau_e^{(k)}$ replaced by $\tau_i^{(k)}$.

For the purpose of model selection and model comparison, we also find the Rao-Blackwellized estimator of the posterior mean of σ_i^2 under Model 2 as

$$\hat{\sigma}_i^{2(\text{RB})} = \frac{1}{G} \sum_{k=1}^G \left[b_i + \frac{1}{2} (Y_i - X_i \beta_i^{(k)})^T (Y_i - X_i \beta_i^{(k)}) \right] \times (a_i + \frac{1}{2} n_i - 1)^{-1}. \quad (14)$$

2.3 Random Error Variances

In Model 2, we assumed unequal error variances for the sampling errors. Kleffe and Rao (1992) used a simple random error variance model to derive the best linear

unbiased predictors for small area means. In this section we extend their model to the multi-level case. We assume random effect models on both regression coefficients β_i and sampling error variances σ_i^2 , which leads to Model 3 given below.

Model 3:

- (i) Same as in Model 2;
(ii) Same as in Model 2;
(iii) Conditional on η and λ , τ_i 's are independent with

$$\tau_i | \eta, \lambda \stackrel{\text{iid}}{\sim} G(\eta, \lambda), \quad (15)$$
where $\tau_i = \sigma_i^{-2}$;
(iv) Marginal prior distributions: $\gamma \sim N_q(0, D)$, $\Omega \sim W_p(\alpha, R)$, $\eta \sim U^+$ and $\lambda \sim U^+$, where U^+ denotes a uniform distribution over a subset of R^+ with large but finite length, D, α and R are known.

Remark 3.1: In Model 3, we assume that τ_i 's are iid gamma random variables with unknown hyperparameters η and λ . Thus we have population models for both regression coefficient β_i and sampling variance σ_i^2 . In Model 1 and Model 2, we considered modelling β_i only and assumed vague proper prior distributions on σ_e^2 or σ_i^2 .

Remark 3.2: Assumption (iii) may not be a good population model for all τ_i 's. Alternatively, we can model τ_i in a more realistic way, as in the case of β_i , by specifying a regression model for the logarithm of τ_i . This may require some auxiliary information related to τ_i . In the data analysis of section 3, however, we simply used $G(\eta, \lambda)$ as the population model for τ_i . Generally it is not easy to model the sampling variances when they are unknown.

The full conditional distributions for Gibbs sampling under Model 3 are given by:

- (i) For $i = 1, \dots, m$,

$$[\beta_i | Y, \tau, \gamma, \Omega, \eta, \lambda] \stackrel{\text{iid}}{\sim} N_p((\tau_i X_i^T X_i + \Omega)^{-1} (\tau_i X_i^T Y_i + \Omega Z_i^T \gamma), (\tau_i X_i^T X_i + \Omega)^{-1})$$
- (ii) For $i = 1, \dots, m$,

$$[\tau_i | Y, \beta, \gamma, \Omega, \eta, \lambda] \stackrel{\text{iid}}{\sim} G \left(\eta + \frac{n_i}{2}, \frac{1}{2} (Y_i - X_i \beta_i)^T (Y_i - X_i \beta_i) + \lambda \right)$$
- (iii)
$$[\gamma | Y, \beta, \tau, \eta, \lambda] \sim N_q \left(\left(\sum_{i=1}^m Z_i^T \Omega Z_i + D^{-1} \right)^{-1} \left(\sum_{i=1}^m Z_i^T \Omega \beta_i \right), \left(\sum_{i=1}^m Z_i^T \Omega Z_i + D^{-1} \right)^{-1} \right)$$
- (iv)
$$[\Omega | Y, \beta, \sigma^2, \gamma, \eta, \lambda] \sim W_p \left(\alpha + m, R + \frac{1}{2} \sum_{i=1}^m (\beta_i - Z_i \gamma)(\beta_i - Z_i \gamma)^T \right)$$

$$(v) \quad [\eta | Y, \beta, \tau, \gamma, \Omega, \lambda] \propto [\Gamma(\eta)]^{-m} \lambda^{mn} (\prod_{i=1}^m \tau_i)^\eta$$

$$(vi) \quad [\lambda | Y, \beta, \tau, \gamma, \Omega, \eta] \sim G(m\eta + 1, \sum_{i=1}^m \tau_i)$$

For Model 3, the posterior estimators of β_i and μ_i have the same forms as those given for Model 2. Under Model 3, the Rao-Blackwellized estimator of the posterior mean of σ_i^2 is given by

$$\hat{\sigma}_i^{2(RB)} = \frac{1}{G} \sum_{k=1}^G [\lambda^{(k)} + \frac{1}{2} (Y_i - X_i \beta_i^{(k)})^T \times (Y_i - X_i \beta_i^{(k)}) (\eta^{(k)} + \frac{1}{2} n_i - 1)^{-1}]. \quad (16)$$

Under Model 3, $[\eta | Y, \beta, \tau, \gamma, \Omega, \lambda]$ is known only up to a multiplicative constant. However, since $[\eta | Y, \beta, \tau, \gamma, \Omega, \lambda]$ is a log-concave function of η (see Appendix A2), adaptive rejection sampling method of Gilks, Best and Tan (1995) can be used in the Gibbs sampler to generate samples from the conditional distribution $[\eta | Y, \beta, \tau, \gamma, \Omega, \lambda]$.

3. DATA ANALYSIS

3.1 Data and Model Description

Following Holt and Moura (1993) and Moura and Holt (1999), we considered the estimation of household income in some counties (small areas) of Brazil. Holt and Moura's original data contains 140 small areas with the sampling units taken from each area by simple random sampling. The hierarchical Bayes method does not require the number of small areas to be large, unlike in the case of EBLUP method, for getting standard errors. Therefore, we used only a small part of the original data set in our data analysis for simple illustration. Our data set contains a subset of 10 small areas with 28 sampling units obtained by simple random sampling in each area.

Let y_{ij} denote the j -th household's income in the i -th small area. There are two unit level auxiliary variables, namely x_1 and x_2 , where x_1 denotes the number of rooms in a household and x_2 denotes the educational attainment of Head of Household. The sampling model is given by

$$y_{ij} = x_{ij}^T \beta_i + e_{ij} = \beta_{0i} + x_{1ij} \beta_{1i} + x_{2ij} \beta_{2i} + e_{ij}, \quad (17)$$

where x_{1ij} denotes the number of rooms in the j -th household of small area i and x_{2ij} denotes the corresponding educational attainment of Head of Household. Values of x_{1ij} and x_{2ij} are centered around their respective overall sample means and e_{ij} is the sampling error variable with its distribution specified by the three error variance models discussed in section 2.

In the sampling model (17), β_i is the random regression coefficient corresponding to the i -th small area and is modelled as

$$\beta_{0i} = \gamma_0 + v_{0i}, \beta_{1i} = \gamma_{10} + \gamma_{11} z_i + v_{1i}, \beta_{2i} = \gamma_{20} + \gamma_{21} z_i + v_{2i},$$

where $\gamma = (\gamma_0, \gamma_{10}, \gamma_{11}, \gamma_{20}, \gamma_{21})^T$ is the unknown vector of fixed regression parameters, $v_i = (v_{0i}, v_{1i}, v_{2i})^T$ is the i -th small area random effect vector distributed as $v_i \sim N_3(0, \Phi)$, and z_i is an area level variable defined as the average number of cars per household in each small area. Value of z_i is also centered around its overall sample mean.

We used the three models discussed in section 2 for our data analysis. Vague proper prior distributions on unknown parameters are specified as follows: $\gamma \sim N_5(0, D)$ where $D = \text{diag}(10^4, 10^4, 10^4, 10^4, 10^4)$, thus $\gamma_0, \gamma_{10}, \gamma_{11}, \gamma_{20}, \gamma_{21}$ are assumed to be independent normal variables with a mean of 0 and a standard deviation of 100, so that a 95% prior interval is around ± 200 , and the prior will be locally uniform over the region supported by the likelihood. Alternatively a uniform prior on a suitably wide interval could be given, such as $U(-200, 200)$. A Wishart prior $W_3(\alpha, R)$ is specified for the inverse covariance matrix $\Omega = \Phi^{-1}$. To represent vague prior knowledge, we have chosen the degrees of freedom α for this distribution to be as small as possible, i.e., $\alpha = 3$, the rank of Ω (Spiegelhalter, Thomas, Best and Gilks 1996). The scale matrix R is specified with diagonal elements equal to 1 and off-diagonal elements equal to 0.001, which represents our prior guess at the order of magnitude of the covariance matrix. For Model 1 and Model 2, a gamma prior $G(0.001, 0.001)$ is assumed for τ_e and τ_i 's. For Model 3, $\tau_i \sim G(\eta, \lambda)$, and η and λ are assumed to be independently distributed as $U(0, 10000)$, i.e., the uniform distribution over a large interval. We anticipate that the vague proper priors on the hyperparameters would approximate the flat priors reasonably well and thus would have minimal effect on the posterior estimation.

We implemented the Gibbs sampler for the three models using the BUGS program (Spiegelhalter *et al.* 1996), aided by CODA Splus function (Best, Cowles and Vines 1996) for assessing convergence. The BUGS program constructs the necessary full conditional distributions and carries out the Gibbs sampling as long as we specify our models using the BUGS language. Priors and initial values of the parameters must be specified in the program. For each model, the Gibbs sampler was first run for a "burn-in" period of 2,000 iterations, then 5,000 more iterations were run and kept for model analysis and estimation.

Our interest is to estimate the small area mean $\mu_i = \bar{X}_i^T \beta_i = \beta_{0i} + \bar{X}_{1i} \beta_{1i} + \bar{X}_{2i} \beta_{2i}$, where \bar{X}_{1i} and \bar{X}_{2i} are the i -th small area population means of the auxiliary variables x_1 and x_2 , respectively. For this, we will first select a model for the data set, then we will present the model-based estimates for the small area means based on the selected model.

3.2 Model Selection

We have proposed three models in section 2 based on different assumptions on sampling variances. To examine

which model fits the data, we first obtained the posterior estimates of the sampling variances under the three models. We also calculated the ordinary least square (OLS) estimates of the sampling variances within each area using only the area-specific data. Table 1 shows the Rao-Blackwellized estimates of the sampling variances under the three models as well as the OLS estimates.

Table 1
Estimated Sampling Error Variances

Area	OLS	Model 1	Model 2	Model 3
1	38.17	76.86	40.18	63.60
2	31.75	76.86	34.24	62.13
3	81.26	76.86	94.77	79.58
4	48.73	76.86	52.01	67.27
5	115.98	76.86	121.65	87.70
6	90.74	76.86	94.35	79.78
7	101.67	76.86	101.67	82.14
8	135.65	76.86	159.94	97.96
9	59.10	76.86	63.37	70.57
10	62.86	76.86	65.72	71.22

From Table 1, the OLS estimates indicate large variations among the ten small areas. Model 1 assumes an equal error variance σ_e^2 for all areas and σ_e^2 is estimated by $\hat{\sigma}_e^{2(RB)} = 76.86$, which is much smaller than the OLS estimates for some areas. Model 2 assumes unequal error variances σ_i^2 across areas. Under Model 2, the estimated error variances $\hat{\sigma}_i^{2(RB)}$ to some extent show the feature of the areas; $\hat{\sigma}_i^{2(RB)}$ are consistent with the pattern of the OLS estimates. The most notable result is $\hat{\sigma}_5^{2(RB)} = 121.65$ and $\hat{\sigma}_8^{2(RB)} = 159.94$, which show that there are larger variations within small areas 5 and 8. Model 3 assumes σ_i^2 's to be random variables distributed as $G(\eta, \lambda)$. Under Model 3, all $\hat{\sigma}_i^{2(RB)}$ tend to be equal to and have moved toward $\hat{\sigma}_e^{2(RB)} = 76.86$. The results in Table 1 suggest that Model 2, the unequal error variance model, could be the best model for our data set. For further investigation, we now present a cross-validation study to select a best fit model.

In order to study how the data support each model, we calculated the cross-validation predictive densities for each data point y_{ij} . The cross-validation density for y_{ij} is the conditional density $f(y_{ij}|Y_{(ij)})$, where $Y_{(ij)}$ denotes all data except y_{ij} . We looked at the value of $f(y_{ij}|Y_{(ij)})$ at the observed data point, the so called conditional predictive ordinate, or CPO, for each of the three models. That is

$$CPO_{ij} = f(y_{ij, obs}|Y_{(ij), obs}),$$

where $y_{ij, obs}$ denotes the observed data point. Since CPOs are nothing but the observed likelihoods, models with larger CPOs provide better fit to the observed data. By using the output from the Gibbs sampler, we can calculate the CPOs for all data points. For example, under Model 1, we have

$$\begin{aligned} f(y_{ij}|Y_{(ij)}) &= \frac{f(Y)}{f(Y_{(ij)})} \\ &= \frac{1}{\int \frac{f(Y_{(ij)}|\beta_r, \sigma_e^2)}{f(Y|\beta_r, \sigma_e^2)} \cdot f(\beta_r, \sigma_e^2|Y) d\beta_r d\sigma_e^2} \\ &= \frac{1}{\int \frac{1}{f(y_{ij}|Y_{(ij)}, \beta_r, \sigma_e^2)} \cdot f(\beta_r, \sigma_e^2|Y) d\beta_r d\sigma_e^2}. \end{aligned}$$

Now noting that the y_{ij} 's are conditionally independent, i.e., $f(y_{ij}|Y_{(ij)}, \beta_r, \sigma_e^2) = f(y_{ij}|\beta_r, \sigma_e^2)$, the CPO values are calculated as

$$\widehat{CPO}_{ij} = \frac{1}{\frac{1}{G} \sum_{k=1}^G \frac{1}{f(y_{ij, obs}|\beta_i^{(k)}, \sigma_e^{2(k)})}}, \quad (18)$$

where $f(y_{ij}|\beta_r, \sigma_e^2)$ is the normal density function given by (3). For Model 2 and Model 3, the CPOs are calculated with $\sigma_e^{2(k)}$ replaced by $\sigma_i^{2(k)}$ in (18). More detailed discussion can be found in Gelfand (1995).

We present a CPO plot for the three models in Figure 1. Clearly Model 2 is the best model among the three, because a majority of CPO values for Model 2 are significantly larger than those for Model 1 and Model 3. Model 3 is slightly better than Model 1 in terms of CPO values. Also there are small CPO values for all three models, which indicate that our model assumptions may not be very well satisfied by our data set.

According to the sampling variance estimates given in the Table 1 and the CPO plot, we conclude that Model 2 is a good model for our data. Therefore, we used Model 2 to find model-based estimates of small area means and associated posterior standard errors.

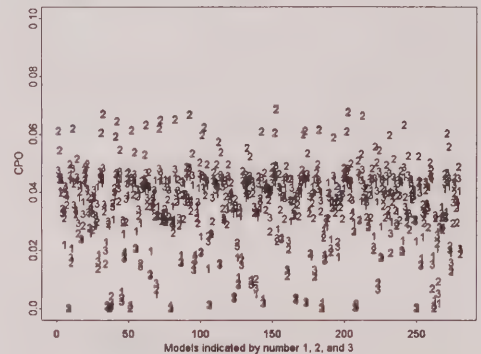


Figure 1. Model selection: CPO comparison plot

3.3 Result of Estimation

We now present the estimates of the small area means based on Model 2 only. Table 2 presents the estimated posterior small area means and the corresponding posterior

standard errors. Our study found that the empirical estimator $\hat{\mu}_i^{(E)}$ and the Rao-Blackwellized estimator $\hat{\mu}_i^{(RB)}$ gave almost the same point estimates, thus we only reported the estimates obtained by using $\hat{\mu}_i^{(RB)}$. For comparison, we also calculated the direct estimates (sample means) and corresponding direct standard errors for the ten areas. It is clear from Table 2 that the model-based estimates are substantially more efficient than the direct estimates. The posterior standard errors are much smaller than the direct standard errors.

Table 2
Estimates of Small Area Means

Area	\bar{y}_i	s.e.	$\hat{\mu}_i^{(RB)}$	s.e.
1	11.08	9.53	10.23	0.81
2	7.91	6.82	9.84	0.85
3	13.48	14.15	13.01	1.08
4	6.53	8.01	10.95	1.11
5	19.52	14.96	17.87	1.57
6	11.21	11.38	10.21	0.93
7	8.72	11.24	9.58	0.97
8	12.81	13.99	10.30	1.19
9	10.18	8.76	11.34	1.01
10	10.01	11.30	9.79	0.87

In order to study the effects of Rao-Blackwellization, we calculated the simulation standard errors of $\hat{\mu}_i^{(E)}$ and $\hat{\mu}_i^{(RB)}$, which are respectively the sample standard errors of $\{\bar{X}_i^T \hat{\beta}_i^{(k)}\}$ and $\{\bar{X}_i^T E[\beta_i | Y, \gamma^{(k)}, \Omega^{(k)}, \tau_e^{(k)}]\}$. Table 3 presents the simulation standard errors. It is clear from Table 3 that the Rao-Blackwellized estimator $\hat{\mu}_i^{(RB)}$ has much smaller simulation standard error than the empirical estimator $\hat{\mu}_i^{(E)}$ for all areas. In all cases the standard error of $\hat{\mu}_i^{(RB)}$ is about 50% to 75% of the standard error of $\hat{\mu}_i^{(E)}$, demonstrating the benefit of Rao-Blackwellization. Thus $\hat{\mu}_i^{(RB)}$ is more stable than $\hat{\mu}_i^{(E)}$ when used to produce point estimates for the posterior means in computational Bayesian analysis. It should be mentioned that the simulation standard error of $\hat{\mu}_i^{(E)}$ is also an estimator of the posterior standard error. Thus the simulation standard error of $\hat{\mu}_i^{(E)}$ in Table 3 is almost identical to the estimated standard error of $\hat{\mu}_i^{(RB)}$ in Table 2.

Table 3
Simulation Standard Errors

Area	$\hat{\mu}_i^{(E)}$	$\hat{\mu}_i^{(RB)}$
1	0.817	0.506
2	0.862	0.498
3	1.090	0.548
4	1.101	0.604
5	1.583	0.878
6	0.930	0.481
7	0.978	0.480
8	1.208	0.842
9	0.997	0.524
10	0.869	0.513

3.4 Sensitivity Analysis

In Model 2, the error variances $\tau_i = \sigma_i^{-2}$ are assumed to be independent with prior distributions $G(a_i, b_i)$ or σ_i^2 with the inverse gamma $IG(a_i, b_i)$, where a_i and b_i are known values chosen to reflect our prior knowledge about σ_i^2 . In practice, it is always difficult to obtain accurate information about the sampling variances. Also, as the number of small areas m increases, the number of variance components σ_i^2 will increase. We are interested in the possible effects caused by the choice of priors on σ_i^2 's; in particular, we would like to evaluate the sensitivity of the posterior means to the choice of priors on the sampling variances σ_i^2 . In our data analysis, a_i and b_i were chosen to be 0.001. Thus we used proper priors with very small parameter values for the variance components to reflect our vague knowledge about σ_i^2 . In order to test the sensitivity of the posterior estimates to the choice of a_i and b_i under Model 2, we set $a_i = b_i$ at six different values, i.e., 0.0001, 0.001, 0.01, 0.1, 1, and 10. Since

$$\{\tau_i | Y, \beta, \gamma, \Omega\} \sim G(a_i + \frac{1}{2}n_i, b_i + \frac{1}{2}(Y_i - X_i\beta_i)^T (Y_i - X_i\beta_i)), \quad (19)$$

the sample effects $n_i/2$ and $(Y_i - X_i\beta_i)^T (Y_i - X_i\beta_i)/2$ dominate the prior information a_i and b_i when a_i and b_i are small. Thus $IG(0.0001, 0.0001)$, $IG(0.001, 0.001)$, and $IG(0.01, 0.01)$ may be viewed as noninformative priors whereas $IG(1, 1)$ and $IG(10, 10)$ may be regarded as informative priors. Table 4 presents posterior means under Model 2 using the different priors on σ_i^2 , and Table 5 presents the corresponding posterior variances.

Table 4
Comparison of Estimated Small Area Means

Small Area	$IG(a_i, b_i), a_i = b_i$					
	0	0	0.01	0.1	1	10
1	10.23	10.23	10.23	10.24	10.25	10.37
2	9.84	9.84	9.84	9.83	9.82	9.62
3	13.00	13.00	13.01	13.01	13.07	13.09
4	10.95	10.95	10.95	10.95	10.94	10.61
5	17.86	17.87	17.85	17.76	17.78	18.27
6	10.21	10.21	10.21	10.21	10.25	10.28
7	9.58	9.58	9.59	9.58	9.63	9.57
8	10.29	10.30	10.30	10.26	10.37	10.86
9	11.34	11.34	11.35	11.32	11.32	11.23
10	9.79	9.79	9.80	9.79	9.82	9.92

It is clear from Table 4 that the small area mean estimates are very stable: they are not sensitive to the choice of a_i and b_i . However, as shown in Table 5, the posterior variances decrease as the priors on σ_i^2 become more informative, and lead to smaller coefficients of variation (CV). This indicates that we can improve estimation results for small areas in terms of CV if we have more prior information on the sampling error variances. In our study, we only considered the case $a_i = b_i$. A more extensive study would involve different combinations of a_i and b_i .

Table 5
Comparison of Estimated Posterior Variances

Small Area	IG(a_i, b_i), $a_i = b_i$					
	0	0	0.01	0.1	1	10
1	0.658	0.658	0.658	0.656	0.653	0.499
2	0.724	0.724	0.724	0.711	0.684	0.462
3	1.167	1.167	1.167	1.161	1.152	0.917
4	1.220	1.220	1.218	1.217	1.202	0.919
5	2.455	2.455	2.454	2.462	2.139	1.335
6	0.871	0.870	0.870	0.830	0.826	0.699
7	0.933	0.933	0.931	0.930	0.914	0.779
8	1.418	1.417	1.418	1.375	1.351	1.337
9	1.015	1.014	1.014	1.011	0.975	0.790
10	0.760	0.760	0.760	0.750	0.745	0.613

Table 6 presents the posterior estimates of σ_i^2 using the different priors on σ_i^2 . As we can see from Table 6, when a_i and b_i are small (≤ 0.01), there is almost no difference among the estimates at all. As a_i and b_i increase, the estimates $\hat{\sigma}_i^{2(RB)}$ become smaller. However, if there is strong prior information on a_i and b_i , for example, $a_i = b_i = 10$, then the posterior estimates of σ_i^2 will be significantly different from the ones under noninformative priors.

Table 6
Comparison of Estimated Sampling Error Variances

Small Area	IG(a_i, b_i), $a_i = b_i$					
	0	0	0.01	0.1	1	10
1	40.09	40.1	40.05	39.64	37.14	22.29
2	34.19	34.18	34.17	33.97	31.74	19.05
3	94.48	94.49	94.42	93.76	86.73	50.60
4	52.08	52.08	52.04	51.63	48.21	28.82
5	121.60	121.70	121.60	121.40	113.70	66.75
6	94.03	94.03	93.83	92.96	87.21	52.90
7	102.30	102.30	102.20	101.40	94.85	57.58
8	160.10	160.00	159.90	159.10	147.60	86.61
9	63.46	63.46	63.38	62.99	58.46	34.85
10	65.88	65.87	65.89	65.40	60.76	36.60

4. CONCLUDING REMARKS

In this paper, we have presented hierarchical Bayes methods for small area estimation, using multi-level models. Clearly it is not easy to provide a suitable model for all small areas with satisfactory results, even if the Markov Chain Monte Carlo (MCMC) Bayesian methods such as the Gibbs sampling enable us to fit the data using Bayesian models of virtually unlimited complexity. The size and homogeneity of the areas and the availability of good auxiliary information will affect the final results. Models which prove suitable in some situations may be unsuitable in others. The hierarchical Bayes method also has some limitations such as the choice of priors on the model parameters and some sampling issues related to the Gibbs

sampling method. Nevertheless, the general hierarchical Bayes methodology is applicable to a wide variety of situations for estimation of small area parameters. Model selection and choice is an important part of the hierarchical Bayes analysis. It is also important to compare the hierarchical Bayes method with other widely used methods in small area estimation, such as empirical Bayes (EB) and empirical best linear unbiased prediction (EBLUP). Work is in progress on extending our work to account for survey design weights, along the lines of You and Rao (1999).

ACKNOWLEDGMENTS

We would like to thank two referees and the Editor for their helpful comments and suggestions. We also would like to thank Professor F. Moura of Federal University of Rio de Janeiro in Brazil for providing the data set used in section 3. This work was partially supported by a research grant from the Natural Sciences and Engineering Research Council of Canada.

APPENDIX

A1:

The Rao-Blackwellized estimator of the posterior variance of β_i is given by:

$$\begin{aligned}
 \hat{V}(\beta_i) &= \frac{1}{G} \sum_{k=1}^G V(\beta_i | Y, \gamma^{(k)}, \Omega^{(k)}, \tau_e^{(k)}) \\
 &\quad + \frac{1}{G} \sum_{k=1}^G [E(\beta_i | Y, \gamma^{(k)}, \Omega^{(k)}, \tau_e^{(k)})]^2 \\
 &\quad - \left[\frac{1}{G} \sum_{k=1}^G E(\beta_i | Y, \gamma^{(k)}, \Omega^{(k)}, \tau_e^{(k)}) \right]^2 \\
 &= \frac{1}{G} \sum_{k=1}^G (\tau_e^{(k)} X_i^T X_i + \Omega^{(k)})^{-1} \\
 &\quad + \frac{1}{G} \sum_{k=1}^G (\tau_e^{(k)} X_i^T X_i + \Omega^{(k)})^{-1} (\tau_e^{(k)} X_i^T Y_i + \Omega^{(k)} Z_i \gamma^{(k)}) \\
 &\quad \times (\tau_e^{(k)} X_i^T Y_i + \Omega^{(k)} Z_i \gamma^{(k)})^T (\tau_e^{(k)} X_i^T X_i + \Omega^{(k)})^{-1} \\
 &\quad - \frac{1}{G^2} \left[\sum_{k=1}^G (\tau_e^{(k)} X_i^T X_i + \Omega^{(k)})^{-1} (\tau_e^{(k)} X_i^T Y_i + \Omega^{(k)} Z_i \gamma^{(k)}) \right] \\
 &\quad \times \left[\sum_{k=1}^G (\tau_e^{(k)} X_i^T X_i + \Omega^{(k)})^{-1} (\tau_e^{(k)} X_i^T Y_i + \Omega^{(k)} Z_i \gamma^{(k)}) \right]^T.
 \end{aligned}$$

A2:

Lemma: $[\eta|Y, \beta, \tau, \gamma, \Omega, \lambda]$ is a log-concave function of η .

Proof: Let $h(\eta) = \log[\eta|Y, \beta, \tau, \gamma, \Omega, \lambda]$. It is enough to show that

$$\frac{\partial^2 h(\eta)}{\partial^2 \eta} \leq 0.$$

Clearly,

$$\frac{\partial h(\eta)}{\partial \eta} = -m \frac{\Gamma'(\eta)}{\Gamma(\eta)} + m \log(\lambda) + \log(\prod_{i=1}^m \tau_i).$$

Let $\psi(\eta) = \Gamma'(\eta)/\Gamma(\eta)$, then we have

$$\frac{\partial^2 h(\eta)}{\partial^2 \eta} = -m\psi'(\eta) \leq 0$$

since $m > 0$ and $\psi'(\eta)$ is positive on $(0, \infty)$ (Temme, 1994, 54-55).

REFERENCES

- BATTESE, G.E., HARTER, R.M., and FULLER, W.A. (1988). An error components model for prediction of county crop area using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.
- BEST, N., COWLES, M.K., and VINES, K. (1996). CODA, *Convergence Diagnosis and Output Analysis Software for Gibbs Sampling Output*, Version 0.30. MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR.
- GELFAND, A.E. (1995). Model determination using sampling-based methods. In *Markov Chain Monte Carlo in Practice* (W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, Eds.), 145-161. London: Chapman and Hall.
- GELFAND, A.E., and SMITH, A.F.M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.
- GELFAND, A.E., and SMITH, A.F.M. (1991). Gibbs sampling for marginal posterior expectations. *Communications In Statistics - Theory and Methods*, 20, 1747-1766.
- GHOSH, M., and RAO, J.N.K. (1994). Small area estimation: An appraisal (with discussion). *Statistical Science*, 9, 55-93.
- GILKS, W.R., BEST, N.G., and TAN, K.K.C. (1995). Adaptive rejection Metropolis sampling within Gibbs sampling. *Journal of Applied Statistics*, 44, 455-472.
- HOBERT, J.P., and CASSELLA, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*, 91, 1461-1473.
- HOLT, D., and MOURA, F. (1993). Small area estimation using multi-level models. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 21-30.
- KLEFFE, J., and RAO, J.N.K. (1992). Estimation of mean square error of empirical best linear unbiased predictors under a random error variance linear model. *Journal of Multivariate Analysis*, 43, 1-15.
- MOURA, F., and HOLT, D. (1999). Small area estimation using multi-level models. *Survey Methodology*, 25, 73-80.
- PRASAD, N.G.N., and RAO, J.N.K. (1990). The estimation of mean squared error of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.
- RAO, J.N.K. (1999). Some recent advances in model-based small area estimation. *Survey Methodology*, 25, 175-186.
- SPIEGELHALTER, D., THOMAS, A., BEST, N., and GILKS, W. (1996). BUGS 0.5, *Bayesian Inference Using Gibbs Sampling Manual*. MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR.
- TEMME, N.M. (1994). *Special Functions: An Introduction to the Classical Functions of Mathematical Physics*. New York: John Wiley.
- YOU, Y., and RAO, J.N.K. (1999). Pseudo hierarchical Bayes small area estimation using sampling weights. *Proceedings of the Survey Methods Section, Statistical Society of Canada*, 117-122.

Double Sampling for Ratio and Regression Estimation With Sub-sampling the Non-respondents

FABIAN C. OKAFOR and HYUNSHIK LEE¹

ABSTRACT

Cochran (1977, p. 374) proposed some ratio and regression estimators of the population mean using the Hansen and Hurwitz (1946) procedure of sub-sampling the non-respondents assuming that the population mean of the auxiliary character is known. For the case where the population mean of the auxiliary character is not known in advance, some double (two-phase) sampling ratio and regression estimators are presented in this article. The relative performances of the proposed estimators are compared with the estimator proposed by Hansen and Hurwitz (1946).

KEY WORDS: Hansen and Hurwitz estimator; Survey cost; Optimum sampling fraction.

1. INTRODUCTION

In many human surveys, information is in most cases not obtained from all the units in the survey even after some call-backs. An estimate obtained from such incomplete data may be misleading especially when the respondents differ from the non-respondents because the estimate can be biased. Hansen and Hurwitz (1946) proposed a technique for adjusting for non-response to address the bias problem. Their idea is to take a sub-sample from the non-respondents to get an estimate for the subpopulation represented by the non-respondents.

Cochran (1977), using Hansen and Hurwitz (1946) procedure, proposed the ratio and regression estimators of the population mean of the study variable in which information on the auxiliary variable is obtained from all the sample units, while some sample units failed to supply information on the study variable. In addition, the population mean of the auxiliary variable is known. In this paper we shall assume that the population mean of the auxiliary variable is not known. We, therefore, use the double sampling method to estimate the mean of the auxiliary variable and then go on to estimate the mean of the study variable in a similar manner as Cochran (1977).

In practice, non-response is often compensated for by weighting adjustment (Oh and Scheuren 1983) or by imputation (Kalton and Karspryck 1986). The procedures used for weighting adjustment and imputation strive for elimination of the bias due to non-response. However, those procedures are based on untenable assumptions on the response mechanism. When the assumed mechanism is wrong, then the resulting estimate can be seriously biased. Moreover, it is difficult to eliminate the bias entirely when non-response is confounded in the sense that the response probability is dependent on the survey character. Rancourt, Lee, and Särndal (1994) provided a partial correction for the

situation. Hansen and Hurwitz's sub-sampling approach does not have this defect although it costs more because of extra work required for sub-sampling the non-respondents. Nonetheless, if the bias problem is serious, the procedure is a viable option to address the problem without resorting to 100 percent response, which can be very expensive.

In the next section, double sampling ratio and regression estimators are considered. Generally, the double sampling procedure is used when it is necessary to make use of auxiliary information to improve the precision of an estimate but the population distribution of the auxiliary information is not known. The first phase sample is used to estimate the population distribution of the auxiliary variable, while the second phase sample is used to obtain the required information on the variable of main interest. The optimum sampling fractions are derived for the estimators for a fixed cost. The performances of the proposed estimators are compared both theoretically and empirically with the Hansen and Hurwitz estimator.

2. THE DOUBLE SAMPLING RATIO AND REGRESSION ESTIMATORS

2.1 Background

To estimate the population mean \bar{X} of the auxiliary variable, a large first phase sample of size n' is selected from N units in the population by simple random sampling without replacement (SRSWOR). A smaller second phase sample of size n is selected from n' by SRSWOR and the character y is measured on it. The ratio estimator of the mean of y is $\bar{y}'_r = (\bar{y}/\bar{x})\bar{x}'$, where \bar{x}' is the sample mean from n' units, \bar{y} and \bar{x} are obtained from the second phase sample if there is no non-response in the second phase sample. If, however, there is non-response in the second phase sample, we may use an estimator obtained from only

¹ Fabian C. Okafor, Dept. of Statistics, University of Nigeria, Nsukka, Nigeria; Hyunshik Lee, formerly Statistics Canada, now Westat, 1650 Research Boulevard, Rockville, Maryland, 20850, U.S.A.

the respondents or take a sub-sample of the non-respondents and re-contact them. The former option is much cheaper than the latter because securing missing information from the non-respondents by re-contact requires usually much more effort and cost. However, it is quite feasible that the non-respondents differ significantly in the main character from the respondents so that a serious bias results. In this situation, sub-sampling of the non-respondents may be beneficial. Hence, we pursue the sub-sampling idea of Hansen and Hurwitz for a double sampling situation. Basically, the estimators proposed here are double sampling version of Cochran (1977, p. 374), that is, double sampling ratio and regression estimators for \bar{Y} adjusted for non-response by using the Hansen and Hurwitz (1946) procedure.

Let's assume that all the n' units supplied information on the auxiliary variable x at the first phase. But let n_1 units supply information on y and n_2 refuse to respond at the second phase. From the n_2 non-respondents, an SRSWOR of m units is selected with the inverse sampling rate k , where $m = n_2/k$, $k > 1$. All the m units respond this time around. This can be applied in a household survey where the household size is used as an auxiliary variable for the estimation of, say, family expenditure. Information can be obtained completely on the family size during the household listing while there may be non-response on the household expenditure.

In the following presentation, we assume that the whole population (denoted by A) is stratified into two strata: one is the stratum (denoted by A_1) of N_1 units, which would respond on the first call at the second phase and the other stratum (denoted by A_2) consists of N_2 units, which would not respond on the first call at the second phase but will respond on the second call. Let the first and second phase samples be denoted by a' and a respectively, and let $a_1 = a \cap A_1$ and $a_2 = a \cap A_2$. The sub-sample of a_2 will be denoted by a_{2m} . Summation over the units in a set s will be denoted by \sum_s .

As a general rule, population parameters are denoted by capital letters except for Greek letters and the sample statistics by corresponding small letters.

2.2 The Double Sampling Ratio Estimator

We define the double sampling ratio estimator as follows:

$$d^* = \frac{\bar{y}^*}{\bar{x}^*} \bar{x}' = r^* \bar{x}' \quad (2.1)$$

where \bar{x}^* and \bar{y}^* are the Hansen-Hurwitz estimators for \bar{X} and \bar{Y} , respectively, and are given by

$$\bar{u}^* = w_1 \bar{u}_1 + w_2 \bar{u}_{2m}, \quad u = x, y. \quad (2.2)$$

According to the general rule, we define $W_j = N_j/N$ and $w_j = n_j/n$, $j = 1$ or 2 . Sample statistics obtained from a_{2m}

are subscripted by "2m", (e.g., $\bar{u}_{2m} = (1/m) \sum_{a_{2m}} u_i$); those from a_1 are subscripted by "1", (e.g., $\bar{u}_1 = (1/n_1) \sum_{a_1} u_i$), and those for the first phase sample a' will be superscripted by a prime (e.g., $\bar{x}' = (1/n') \sum_{a'} x_i$).

A large sample first order approximation to the variance of d^* , obtained by using the Taylor linearization, is given by

$$V(d^*) \cong \left(\frac{1}{n'} - \frac{1}{N} \right) S_y^2 + \left(\frac{1}{n} - \frac{1}{n'} \right) S_r^2 + \frac{W_2(k-1)}{n} S_{2r}^2 \quad (2.3)$$

where,

$$S_r^2 = S_y^2 + R^2 S_x^2 - 2RS_{xy}, \quad (2.4)$$

$$S_{2r}^2 = S_{2y}^2 + R^2 S_{2x}^2 - 2RS_{2xy},$$

R is the population ratio of \bar{Y} to \bar{X} . S_u^2 and S_{2u}^2 are, respectively, the variance for the whole population and the population variance for the stratum of non-respondents of the variable u . S_{xy} and S_{2xy} are the covariances for the whole population and the population of non-respondents respectively.

The variance of d^* can be approximately estimated by

$$v(d^*) = \left(\frac{1}{n'} - \frac{1}{N} \right) \hat{S}_y^2 + \left(\frac{1}{n} - \frac{1}{n'} \right) \hat{S}_r^2 + \frac{w_2(k-1)}{n} \hat{S}_{2r}^2 \quad (2.5)$$

where,

$$\hat{S}_y^2 = \frac{1}{n-1} \left\{ \sum_{a_1} y_i^2 + k \sum_{a_{2m}} y_i^2 - n\bar{y}^2 + w_2(k-1)s_{2my}^2 \right\},$$

$$\hat{S}_r^2 = \frac{1}{n-1} \left\{ \sum_{a_1} (y_i - r^* x_i)^2 + k \sum_{a_{2m}} (y_i - r^* x_i)^2 \right\} \text{ and}$$

$$\hat{S}_{2r}^2 = \frac{1}{m-1} \sum_{a_{2m}} (y_i - r^* x_i)^2. \quad (2.6)$$

Note that \hat{S}_y^2 is an unbiased estimator of S_y^2 . It seems natural to use \hat{S}_r^2 to estimate S_r^2 since the expression obtained from \hat{S}_r^2 by replacing r^* with R is a consistent estimator of S_r^2 . The same argument can be used to justify the use of \hat{S}_{2r}^2 .

An alternative estimator of $V(d^*)$ can be obtained by replacing \hat{S}_r^2 and \hat{S}_{2r}^2 with

$$\tilde{S}_r^2 = \hat{S}_y^2 + r^{*2} s_{x'}^2 - 2r^* s_{xy}^*, \text{ and}$$

$$\tilde{S}_{2r}^2 = s_{2my}^2 + r^{*2} s_{2x}^2 - 2r^* s_{2mxy}, \quad (2.7)$$

respectively, in (2.5), where,

$$\begin{aligned}s_x'^2 &= \frac{1}{n' - 1} \sum_{a'} (x_i - \bar{x}')^2, \\ s_{2my}^2 &= \frac{1}{m - 1} \sum_{a_{2m}} (y_i - \bar{y}_{2m})^2, \\ s_{2x}^2 &= \frac{1}{n_2 - 1} \sum_{a_2} (x_i - \bar{x}_2)^2, \\ s_{2mxy} &= \frac{1}{m - 1} \left(\sum_{a_{2m}} x_i y_i - m \bar{x}_{2m} \bar{y}_{2m} \right)\end{aligned}$$

and s_{xy}^* is as in (2.9). This alternative estimator is likely to have a smaller variance than the estimator in (2.5) since the estimators $s_x'^2$ and s_{2x}^2 are based on larger samples and therefore more precise.

2.3 The Double Sampling Regression Estimator

We define the regression estimator by

$$t^* = \bar{y}^* + \hat{\beta}^* (\bar{x}' - \bar{x}^*) \quad (2.8)$$

where $\hat{\beta}^*$ is an estimator of $\beta = S_{xy}/S_x^2$. There could be several choices for $\hat{\beta}^*$, but a natural choice would be given by $\hat{\beta}^* = s_{xy}^*/s_x'^2$, where

$$\begin{aligned}s_{xy}^* &= \frac{1}{n - 1} \left(\sum_{a_1} x_i y_i + k \sum_{a_{2m}} x_i y_i - n \bar{x} \bar{y} \right) \text{ and} \\ s_x'^2 &= \frac{1}{n - 1} \left(\sum_{a_1} x_i^2 + k \sum_{a_{2m}} x_i^2 - n \bar{x} \bar{x} \right).\end{aligned} \quad (2.9)$$

It is easy to show that s_{xy}^* and $s_x'^2$ are unbiased for S_{xy} and S_x^2 respectively. An approximate variance of t^* is given as

$$\begin{aligned}V(t^*) &= \left(\frac{1}{n'} - \frac{1}{N} \right) S_y^2 + \left(\frac{1}{n} - \frac{1}{n'} \right) S_l^2 \\ &\quad + \frac{W_2(k - 1)}{n} S_{2l}^2\end{aligned} \quad (2.10)$$

where S_l^2 and S_{2l}^2 are obtained from (2.4) by replacing R with β .

To estimate $V(t^*)$ we can use the following formula:

$$\begin{aligned}v(t^*) &= \left(\frac{1}{n'} - \frac{1}{N} \right) \hat{S}_y^2 + \left(\frac{1}{n} - \frac{1}{n'} \right) \hat{S}_l^2 \\ &\quad + \frac{w_2(k - 1)}{n} \hat{S}_{2l}^2\end{aligned} \quad (2.11)$$

where,

$$\begin{aligned}\hat{S}_l^2 &= \frac{1}{n - 1} \left\{ \sum_{a_1} (y_i - y_i^*)^2 + k \sum_{a_{2m}} (y_i - y_i^*)^2 \right\}, \\ \hat{S}_{2l}^2 &= \frac{1}{m - 1} \sum_{a_{2m}} (y_i - y_i^*)^2 \text{ and} \\ y_i^* &= \bar{y}_i - \hat{\beta}^* (x_i - \bar{x}^*).\end{aligned} \quad (2.12)$$

Like (2.7), a slightly improved estimator of $V(t^*)$ can be obtained by using

$$\begin{aligned}\tilde{S}_l^2 &= \hat{S}_y^2 + \hat{\beta}^{*2} s_x'^2 - 2 \hat{\beta}^* s_{xy}^* \text{ and} \\ \tilde{S}_{2l}^2 &= s_{2my}^2 + \hat{\beta}^{*2} s_{2x}^2 - 2 \hat{\beta}^* s_{2mxy}.\end{aligned} \quad (2.13)$$

3. CHOICE OF SAMPLING FRACTIONS

We shall now deduce the optimum k , n , and n' that minimize the variances of the proposed estimators for a specified cost, or that minimize the cost for a specified variance.

Let's consider a cost function for d^* given by

$$C = c'n' + cn + c_1 n_1 + c_2 m \quad (3.1)$$

where the c 's are the costs per unit defined as follows:

- c' : the unit cost associated with the first phase sample, a' ;
- c : the unit cost of the first attempt on y with the second phase sample, a ;
- c_1 : the unit cost for processing the respondent data on y at the first attempt in a_1 ;
- c_2 : the unit cost associated with the sub-sample, a_{2m} of a_2 .

Since the value of n_1 is not known until the first attempt is made, the expected cost will be used in the minimization. The expected cost is given by

$$E(C) = C^* = c'n' + \left(c + c_1 W_1 + \frac{c_2 W_2}{k} \right) n. \quad (3.2)$$

The optimum values of k , n , and n' that minimize the variance of d^* for a fixed expected cost C^* are obtained by using Lagrange multiplier. The optimum values thus obtained are:

$$k_o = \sqrt{\frac{c_2(S_r^2 - W_2 S_{2r}^2)}{S_{2r}^2(c + c_1 W_1)}} \quad (3.3)$$

$$n_o = \frac{C^* \sqrt{A}}{D \sqrt{G}} \quad \text{and} \quad n_o' = \frac{C^* \sqrt{S_y^2 - S_r^2}}{D \sqrt{c'}} \quad (3.3)$$

where

$$A = S_r^2 + W_2(k_o - 1)S_{2r}^2,$$

$$G = c + c_1 W_1 + \frac{c_2 W_2}{k_o} \quad \text{and}$$

$$D = \sqrt{(S_y^2 - S_r^2)c'} + \sqrt{AG}.$$

If we let $\gamma = c_2/(c + c_1 W_1)$, $\delta = S_r^2/S_{2r}^2$ and $\xi = S_y^2/S_r^2$, then we have

$$k_o = \sqrt{\gamma(\delta - W_2)},$$

$$n_o = \frac{C^* \sqrt{1 + W_2(k_o - 1)/\delta}}{\sqrt{Gc'(\xi - 1)} + G\sqrt{1 + W_2(k_o - 1)/\delta}} \quad \text{and}$$

$$n_o' = \frac{C^* \sqrt{\xi - 1}}{c' \sqrt{\xi - 1} + \sqrt{Gc' \{1 + W_2(k_o - 1)/\delta\}}}. \quad (3.4)$$

The optimum values n_o and n_o' are proportional to the expected cost, C^* . To get the optimum values of k , n , and n' that, minimize $V(t^*)$ we simply substitute S_r^2 and S_{2r}^2 in the above expression in (3.3) with S_l^2 and S_{2l}^2 , respectively. Table 1 shows optimum values of k_o , n_o , and n_o' for given parameters.

Table 1
Optimum Values of k_o , n_o , and n_o'

C^*	c'	c	c_1	c_2	δ	ξ	W_2	γ	k_o	G	n_o	n_o'
200	0.1	0.5	1	2	1	2	0.3	1.67	1.08	1.76	92	382
200	0.1	0.5	1	2	1	4	0.3	1.67	1.08	1.76	81	580
200	0.1	0.5	1	2	2	2	0.3	1.67	1.68	1.56	104	389
200	0.1	0.5	1	2	2	4	0.3	1.67	1.68	1.56	91	590
200	0.1	0.5	1	4	1	2	0.3	3.33	1.52	1.99	83	345
200	0.1	0.5	1	4	1	4	0.3	3.33	1.52	1.99	74	531
200	0.1	0.5	1	4	2	2	0.3	3.33	2.38	1.70	96	361
200	0.1	0.5	1	4	2	4	0.3	3.33	2.38	1.70	85	553
200	0.5	0.5	1	2	1	2	0.3	1.67	1.08	1.76	85	250
200	0.5	0.5	1	2	1	4	0.3	1.67	1.08	1.76	72	366
200	0.5	0.5	1	2	2	2	0.3	1.67	1.68	1.56	96	255
200	0.5	0.5	1	2	2	4	0.3	1.67	1.68	1.56	80	372
200	0.5	0.5	1	4	1	2	0.3	3.33	1.52	1.99	78	228
200	0.5	0.5	1	4	1	4	0.3	3.33	1.52	1.99	67	338
200	0.5	0.5	1	4	2	2	0.3	3.33	2.38	1.70	89	238
200	0.5	0.5	1	4	2	4	0.3	3.33	2.38	1.70	76	351

4. COMPARISON OF THE ESTIMATORS

In this section, the theoretical comparison of the performances of the proposed estimators with respect to the Hansen and Hurwitz (1946) estimator is made first without taking the cost into consideration and then with taking it into account.

4.1 Without Considering the Cost

The variance of the Hansen-Hurwitz estimator is

$$V(\bar{y}^*) = \left(\frac{1}{n} - \frac{1}{N} \right) S_y^2 + \frac{W_2(k-1)}{n} S_{2y}^2 \quad (4.1)$$

where \bar{y}^* is defined as in (2.2).

$$V(\bar{y}^*) - V(d^*) = \left(\frac{1}{n} - \frac{1}{n'} \right) \left(2RS_{xy} - R^2 S_x^2 \right) + \frac{W_2(k-1)}{n} \left(2RS_{2xy} - R^2 S_{2x}^2 \right). \quad (4.2)$$

This is positive (i.e., d^* is more efficient than \bar{y}^*) if $R < 2\beta$ and $R < 2\beta_2$, where $\beta_2 = S_{2xy}/S_{2x}^2$. On the other hand, we have

$$V(\bar{y}^*) - V(t^*) = \left(\frac{1}{n} - \frac{1}{n'} \right) \frac{S_{xy}^2}{S_x^2} + \frac{W_2(k-1)}{n} \beta S_{2x}^2 (2\beta_2 - \beta). \quad (4.3)$$

Therefore, t^* is more efficient than the Hansen-Hurwitz estimator if (4.3) is positive. One particular condition under which this can occur is that $\beta_2 \geq \beta/2$ with $\beta \geq 0$. The conditions we discuss here are sufficient and thus, d^* or t^* can be more efficient than \bar{y}^* under more relaxed conditions.

4.2 Considering the Cost

We shall now compare the proposed estimators with the Hansen-Hurwitz estimator (\bar{y}^*) making use of the cost function given in section 3.

For the estimator \bar{y}^* , if a straight random sample is taken (without using double sampling procedure) for y , the optimum sample size for an expected cost,

$$C^* = \left(c + c_1 W_1 + \frac{c_2 W_2}{k} \right) n,$$

similar to the one in (3.2) can be obtained by the same technique (i.e., Lagrange multiplier) used in section 3 as follows:

$$n_{oHH} = \frac{C^*}{c + c_1 W_1 + c_2 W_2 / k_{oHH}} \text{ and}$$

$$k_{oHH} = \sqrt{\frac{c_2(S_y^2 - W_2 S_{2y}^2)}{S_{2y}^2(c + c_1 W_1)}}. \quad (4.4)$$

Then the optimum variance of the Hansen-Hurwitz estimator becomes

$$V(\bar{y}^*) = \left(\frac{1}{n_{oHH}} - \frac{1}{N} \right) S_y^2 + \frac{W_2(k_{oHH} - 1)}{n_{oHH}} S_{2y}^2. \quad (4.5)$$

If we compare this with $V(d^*)$ with the optimum choices of k , n , and n' in (3.3), then the condition that d^* will be more precise than \bar{y}^* is given by

$$2\rho - Rh > \frac{1}{1 - \theta_1} \times \left\{ \frac{1}{\beta h} (1 - \theta_2 + Q_y - \theta_2 Q_{HHy}) - h Q_x (2\beta_2 - R) \right\} \quad (4.6)$$

where

$$h = \frac{S_x}{S_y}, \theta_1 = \frac{n_o}{n'_o}, \theta_2 = \frac{n_o}{n_{oHH}}, Q_{HHy} = \frac{W_2(k_{oHH} - 1)S_{2y}^2}{S_y^2},$$

$$Q_u = \frac{W_2(k_o - 1)S_{2u}^2}{S_u^2}, u = x, y,$$

and ρ is the correlation coefficient between x and y .

We can obtain a similar comparison between \bar{y}^* and t^* . That is, t^* is more efficient than \bar{y}^* if

$$2\rho - \beta h > \frac{1}{1 - \theta_1} \times \left\{ \frac{1}{\beta h} (1 - \theta_2 + Q_y - \theta_2 Q_{HHy}) - h Q_x (2\beta_2 - \beta) \right\}. \quad (4.7)$$

4.3 Empirical Comparison of the Proposed Estimators

The relative efficiencies of the estimators d^* and t^* with respect to \bar{y}^* are compared using an artificially generated population. The parameters of the population are:

$$R = 1.92, \beta = 1.52, \rho = 0.85, R_2 = 1.88, \beta_2 = 1.47,$$

$$\rho_2 = 0.83, N = 1,000, N_2 = 302, S_x^2 = 766.54,$$

$$S_y^2 = 2426.82, S_{xy} = 1164.08, S_{2x}^2 = 433.63,$$

$$S_{2y}^2 = 1350.05, \text{ and } S_{2xy} = 638.32.$$

The relative efficiencies of d^* and t^* are presented in Table 2. Note that R is substantially different from β , which means that the regression line does not pass through the origin. Under this population, the regression estimator t^* is more efficient than the ratio estimator d^* . We notice also that the optimum initial sample size, n'_o for t^* is more than for the estimator d^* . The reverse is the case for the optimum second phase sample size n_o . This is so because the regression estimator can be more precise with a smaller second phase sample size so that it allows to allocate more to the first phase sample. Finally the optimum inverse sampling rate k_o is practically the same for the two estimators.

When the linear regression line passes through the origin, the advantage of t^* over d^* disappears, as expected and confirmed in another empirical comparison not shown here.

Table 2
The Relative Efficiencies of d^* and t^* with
Respect to \bar{y}^* ($C^* = 200$, $c = 0.5$, $c_1 = 1$)

c'	c_2	k_{oHH}	n_{oHH}	k_o	n_o	n'_o	Efficiency
d^*							
0.1	2	1.58	127	1.46	92	514	1.85
0.1	4	2.23	115	2.06	85	477	1.91
0.3	2	1.58	127	1.46	78	250	1.23
0.3	4	2.23	115	2.06	73	234	1.32
t^*							
0.1	2	1.58	127	1.47	89	563	2.11
0.1	4	2.23	115	2.08	83	523	2.19
0.3	2	1.58	127	1.47	74	269	1.34
0.3	4	2.23	115	2.08	70	253	1.45

5. CONCLUSIONS

We proposed ratio and regression estimators based on the double sampling procedure when there is non-response on the main character and the population mean of the auxiliary variable is not known. The potentially serious non-response bias is eliminated by sub-sampling the non-respondents as in the Hansen and Hurwitz procedure (1946). We derived optimum sample sizes for a given set of unit costs and compared theoretically and empirically the performance of our estimators with that of the Hansen and Hurwitz estimator.

When there is a strong linear relationship between the main character and the auxiliary character and the auxiliary data can be collected cheaply with a large sample size, our estimators are substantially superior to the Hansen and Hurwitz estimator. Our procedure can be useful when there is a serious concern about the nonresponse bias that is difficult to handle with the usual weighting adjustment or imputation.

ACKNOWLEDGEMENT

We are grateful to the referees and associate editors for their comments that helped to improve our paper.

REFERENCES

- COCHRAN, W.G. (1977). *Sampling Techniques*, 3rd edition. New York: John Wiley & Sons.
- HANSEN, M. H., and HURWITZ, W. N. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association*, 41, 517-529.
- KALTON, G., and KASPRZYK, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1-16.
- OH, H.L., and SCHEUREN, F.J. (1983). Chapter 13. Weighting adjustment for unit non-response. In *Incomplete Data in Sample Surveys*, (I. Olkin, W.G. Madow, and D.B. Rubin, Eds.). Theory and Bibliographies, 2, 143-184. New York: Academic Press.
- RANCOURT, E., LEE, H., and SÄRNDAL, C.-E. (1994). Bias corrections for survey estimates from data with ratio imputed values for confounded nonresponse. *Survey Methodology*, 20, 137-147.

Modeling Interviewer Effects in Panel Surveys: An Application

JAN PICKERY and GEERT LOOSVELDT¹

ABSTRACT

In this paper we will combine two applications of multilevel models. The multilevel model is suitable to analyze interviewer effects on survey data. It can also be used to analyze longitudinal – “repeated measurements” – data. We will analyze a data quality indicator of panel data that come from the Belgian Election Studies. These panel data consist of only two waves. The respondents that cooperated twice are for the most part not interviewed by the same interviewers. This results in a complex data structure with measurements nested in respondents, and respondents nested in interviewers, but without an overall hierarchical nesting structure: cross-classification. This complicated data structure will be analyzed in two different ways: an analysis of all respondents and an analysis of only those who are interviewed twice by the same interviewer. The results of these different analyses will be compared. We conclude that the multilevel cross-classified model is a very flexible and useful tool to analyze interviewer effects in panel surveys.

KEY WORDS: Multilevel models; Cross-classifications; Panel surveys; Interviewer effects; Don't know answer.

1. INTRODUCTION

In this paper we analyze the effect of respondent and interviewer characteristics on the number of “don't know” answers in two waves of the panel survey from the Belgian Election Studies. We use different multilevel models for a subset of the dataset and for the entire dataset. The main purpose of the article is to illustrate how interviewer effects in a panel survey can be analyzed using multilevel models.

A multilevel or hierarchical model is an appropriate tool to analyze data with nested structures, *e.g.*, pupils nested in schools or patients in hospitals. A multilevel model can include variables of the different levels of nesting, but it also takes account of the variability associated with each level. The typical quality of the models is not the functional form relating the variables of the different levels, but rather a more sophisticated treatment of the error structure (DiPrete and Forristal 1994, 334). In education research for instance a multilevel model can account for variation between schools and variation between pupils. Moreover the model tries to replace this variance attributed to both levels by variables of either level. These models are described in various textbooks like Bryk and Raudenbush (1992), Goldstein (1995), Kreft and de Leeuw (1998) and Snijders and Bosker (1999).

Multilevel or hierarchical models also offer the best possibilities to analyze interviewer effects on survey data (Hox 1994). A hierarchical model is the best tool to tackle the “respondents nested within interviewers” – design. Other statistical techniques require mutual independence of interviewer and respondent characteristics, which is – most of the time – not the case because of the hierarchical structure of the data. In a multilevel model both the regression

coefficients and the variance components are conditional on the explanatory variables in the model, which is a useful property if there is no complete orthogonalization of interviewer and respondent variables (Hox 1994, 307). When respondents are not randomly assigned to interviewers, respondent and interviewer characteristics can become confounded since respondents from a specific area will most likely be interviewed by interviewers from the same area. In such a situation, if the relevant respondent variables are put in the multilevel model, interviewers are equalized by statistical means. For that reason the assumptions of an analysis of interviewer effects with a multilevel model are more realistic than those of an ANOVA or ANCOVA. Furthermore the hierarchical model allows estimation of both the interviewer variance and the effects of explanatory variables measured at the interviewer and the respondent model. This possibility of replacing variance attributed to respondents/interviewers with the effects of respondent/interviewer characteristics allows for wider generalizations.

The multilevel model can also fruitfully be used to analyze longitudinal – “repeated measurements” – data (see *e.g.*, Goldstein 1995, 87-95; Snijders 1996 and Yang and Goldstein 1996). There are alternatives to analyze the “measurements nested in individuals” – design, but multilevel analysis has some clear advantages. Using a hierarchical model, it is feasible to handle unbalanced designs – not all individuals have the same number of measurements – and quite easy to incorporate changing covariates. Besides, the model allows more nesting levels. The individuals can be nested in another higher level unit.

We will analyze respondent and interviewer effects on the number of “don't know” answers on a series of questions regarding political parties in a panel survey. We have

¹ Jan Pickery and Geert Loosveldt, Department of Sociology, University of Leuven, E. Van Evenstraat 2B, 3000 Leuven, Belgium. E-mail: jan.pickery@soc.kuleuven.ac.be; geert.loosveldt@soc.kuleuven.ac.be.

measurements (wave 1 and wave 2) nested in respondents (the longitudinal design) and respondents nested in interviewers. Our panel data consist of two waves. During the second wave the respondents are for the most part not interviewed by the same interviewers. The purely hierarchical nesting has broken down and a more complex data structure is the result. To handle this data structure it is necessary to conceive the measurements as being nested into two different classification structures: measurements nested in respondents and in interviewers. This is called a cross-classified design, because the nesting of the levels is not purely hierarchical.

In this article we'll start with a simple data structure and the appropriate model. Afterwards the model will become more complex. We'll perform two analyses. In our first analysis we work only with the respondents who are interviewed twice by the same interviewers. Afterwards we will analyze all the respondents, including those who were interviewed only once. In the first analysis the purely hierarchical structure remains intact. The model is a "simple" three level one: measurements nested in respondents nested in interviewers. Analysis 2 sets up a cross-classified model. In that model the measurements are classified by respondents and interviewers.

The next section reflects on the nature of our dependent variable, the "don't know" answer, and the way to analyze it. In the third section we'll describe our data in detail to clarify the complex structure. The following section (4) treats in brief the different models that we will combine. Section 5 presents the variables in our analysis. In sections 6 and 7 we discuss the setup of our 2 different models and report the results of the analyses. Section 8 concludes the article.

2. ANSWERING "DON'T KNOW"

It has become generally acknowledged that the use of a "don't know" or a "no opinion" filter increases the proportion of respondents who give this answer, and that the increase itself is a function of the nature of the filter used (Schuman and Presser 1981, 143). Krosnick argues that answering "don't know" is one form of satisficing. Satisficing occurs when a respondent is not motivated to expend the mental effort necessary to generate optimal answers. A "no opinion" answer is an acceptable answer but it is the result of a "weak" cognitive process. Satisficing is a function of task difficulty, and the respondent's lack of knowledge, ability and motivation. This theoretical reasoning is consistent with the finding that offering a "don't know" response option increases the proportion of respondents who select it, particularly among respondents with little formal education and people who consider an issue to be less personally important. (Krosnick 1991). Following this argumentation, answering "don't know" is mainly explained by respondent characteristics that can be

related to the cognitive aspect of answering questions. Previous research points us to the following characteristics of interest: education (*e.g.*, Sudman and Bradburn 1974), age (see *e.g.*, Groves 1989, 441-443), sex (*e.g.*, Hox, de Leeuw and Kreft 1991), and a measure of involvement or interest in the subject (*e.g.*, Groves 1989, 419).

However answering questions is not only a cognitive process of the respondent but it is also a communicative process (Schwarz and Sudman 1995). Within this process the interviewer plays an important role. There is a lot of literature about the interviewer as a source of survey measurement error (Groves 1989). The main idea is that interviewers are not "neutral" collectors of data but that they can influence the respondents' answers. Item non-response too is subject to interviewer effects as has been shown long ago by *e.g.*, Hanson and Marks (1958) and Bailar, Bailey and Stevens (1977). A social scientist interested in the explanation of "don't know" answers should therefore include respondents and interviewers in the analysis. The number of "don't know" answers will be the dependent variable of our analyses.

3. DESCRIPTION OF THE DATA STRUCTURE

After the 1991 General Election in Belgium a national survey was set up in which 4,544 face to face interviews were conducted in the three Regions in the early months of 1992. A two-stage self-weighting sample (see *e.g.*, Särndal, Swensson and Wretman 1992, 141-144) was used. The sample was representative for the population of 18-74 years old (ISPO/PIOP 1995). In this article we will use the data from the Flemish region, which cover 2,691 Flemish respondents, interviewed by 163 interviewers (Carton, Swyngedouw, Billiet and Beerten 1993). After the 1995 Elections a similar survey was set up. Due to budgetary constraints the sample had to be smaller for the second wave. So the 2,691 respondents were used as a group to sample from and, in second order, there had to be new respondents to compensate for the aging of the youngest cohort from 1991. Finally 2,099 respondents were interviewed by 167 interviewers. This sample contained 1,762 panel respondents and 337 new respondents (see Beerten, Billiet, Carton and Swyngedouw 1997 for a detailed technical report of the sample plan). Only 55 of the interviewers of the first wave collaborated again. So there were 112 new interviewers in the second wave.

This gives us a dataset with 3028 respondents (2,691 + 337) and 275 interviewers (163 + 112). For 1,762 respondents we have a measurement in both waves, for the rest (1266) there is only one measurement. The structure of the dataset can be represented in a table similar to table 1 (see also Goldstein 1995, 114). Each x in the table reflects an observation. The complete dataset contains 4,790 observations ($(1,762 \times 2) + 1,266$). Each type of respondent in the table represents a possible occurrence in the dataset.

Table 1
A Representation of the Dataset

	Respondent Type 1		Respondent Type 2		Respondent Type 3		Respondent Type 4		N (Interviewers)
Wave	1	2	1	2	1	2	1	2	
Interviewer Type A	x	x	x					x	47
Interviewer Type B				x		x			8
Interviewer Type C			x			x			108
Interviewer Type D				x				x	112
N (Respondents)	374		1388		929		337		

This table illustrates that we have three kinds of respondents: panel respondents who are interviewed twice by the same interviewer (Type 1), respondents who cooperated twice but were interviewed by different interviewers (Type 2) and respondents who are only interviewed once (Type 3 and 4). Our 2 different analyses are based on these different types of respondents. In Analysis 1 we'll look at the respondents whose situation corresponds with that of Type 1. Only 374 Respondents satisfy this condition. They were interviewed twice by the same interviewer. Analysis 2 takes all the 3028 respondents into account (Respondents 1 to 4 of the table).

Furthermore the table shows that we can also discern different interviewers: interviewers who collaborated twice (Type A and B) and interviewers who collaborated only the first wave (Type C) or only the second wave (Type D). The interviewers of Type B collaborated in both waves, but never interviewed the same respondents twice (unlike the interviewers of Type A).

To analyze this complex data structure we will combine three different models that are presented in the next section.

4. A SHORT DESCRIPTION OF THE DIFFERENT MULTILEVEL MODELS USED IN THE ANALYSES

4.1 The General Multilevel Model

The first model we need is the general multilevel model, which has the following form:

$$Y_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + e_{ij}, \quad (1)$$

$$\beta_{0j} = \beta_0 + u_{0j} \text{ and } \beta_{1j} = \beta_1 + u_{1j} \quad (2)$$

or

$$\beta_{0j} = \beta_0 + \gamma_{01}z_{1j} + u_{0j} \text{ and } \beta_{1j} = \beta_1 + \gamma_{11}z_{1j} + u_{1j}. \quad (3)$$

Subscript i refers to the level 1 unit and subscript j to the level unit 2. In our situation level 1 indicates the respondent and level 2 the interviewer. So the response variable Y of

respondent i , interviewed by interviewer j is dependent on the x variable of that respondent. This relationship looks like an ordinary regression model but the parameters of the model are interviewer specific. The β 's differ from interviewer to interviewer. For each β , there is an interviewer residual (u_{0j} or u_{1j}). The β 's can also be made dependent on higher level variables (interviewer characteristics), allowing for generalization across interviewers. We have one second level variable z . Substituting (3) into (1) results in the following overall model:

$$Y_{ij} = \beta_0 + \beta_1x_{1ij} + \gamma_{01}z_{1j} + \gamma_{11}z_{1j}x_{1ij} + u_{1j}x_{1ij} + u_{0j} + e_{ij}. \quad (4)$$

Of course more x and z variables can be included in these relationships. We assume that the residuals u_{0j} , u_{1j} and e_{ij} have means 0 given the values of the explanatory variables z and x . Furthermore it is assumed that the level 1 residuals (e_{ij}) are independent. The level 2 residuals (u_{0j} and u_{1j}) are assumed to be independent from e_{ij} and to have a joint multivariate normal distribution with covariance matrix Ω . They don't have to be independent from each other. Usually they are correlated.

4.2 The Multilevel Model for Longitudinal Analysis

The second model we need is the longitudinal multilevel model. In an analysis of a "repeated measurements" – design with a hierarchical model, the measurements are considered to be the first level and the individual the second. Most of the time the individual units will be persons, but of course they can be other units, like *e.g.*, schools or countries. In our analysis the individuals are the respondents. The analysis tries to estimate a growth curve on the base of the different measurements and to compare differences in curves given individual characteristics. Each observed value is made conditional upon the time of measurement – which can be a measure of time, but also age – and possible transformations of this measurement. Usually the curve is assumed to be a polynomial, which has the following form:

$$Y_{it} = \pi_{0i} + \pi_{1i}t + \pi_{2i}t^2 + \dots + \pi_{ki}t^k + e_{it}. \quad (5)$$

Y_{it} is the observed value for respondent i on moment t , t can be time of measurement or age. π_{hi} ($h = 0 \dots k$) are the trajectory parameters or growth parameters for subject i , k is the degree of the polynomial. In a simple case k has the value 1 and then there is a linear curve. If there are m moments of measurement, a polynomial with degree of $m - 1$ will result in an exact reproduction of the curve. Of course it is more interesting to use a polynomial with a lower degree if that yields a satisfactory reproduction of the curve. You can test whether the model with degree $k + 1$ results in a significant improvement compared to the model with degree k .

The growth parameters have also a subscript for the individual (respondent). The model states that these parameters differ from individual to individual. The second part of the model defines these parameters:

$$\pi_{0i} = \pi_0 + r_{0i} \quad (6)$$

or

$$\pi_{0i} = \pi_0 + \beta_{01}x_{1i} + r_{0i}. \quad (7)$$

The individual parameter equals a general parameter (π_0) + an individual residual (r_{0i}). By the inclusion of individual characteristics (x) it may be possible to reduce the individual specific part, thus generalizing across respondents. In line with (1) and (2) we chose x to denote the individual (respondent) characteristics. But it is worth mentioning that in this model the x variables are higher level (level 2) variables. The individual characteristics can be fixed (the same for all moments of measurement) or varying.

4.3 Cross-Classified Models

The third model we will use is the cross-classified model. Not all data structures are purely hierarchical. Units may be classified along more than one dimension (see Goldstein 1995, 113-116). For example students can be classified by the school they go to and by the neighborhood they live in. In our example measurements are classified by respondents and by interviewers. A cross-classified model has the following form (subscripts j_1 and j_2 refer to the 2 different classification structures):

$$Y_{ij_1j_2} = \beta_{0j_1j_2} + \beta_{1j_1j_2}x_{1ij_1j_2} + e_{ij_1j_2}, \quad (8)$$

$$\beta_{0j_1j_2} = \beta_0 + u_{0j_1} + u_{0j_2} \text{ and } \beta_{1j_1j_2} = \beta_{1j_1j_2} + u_{1j_1} + u_{1j_2}. \quad (9)$$

Equation (9) can be reformulated the same way as equation (3).

$Y_{ij_1j_2}$ is the observed value for individual i , classified by j_1 and j_2 . In our case: the observed value for measurement i on respondent j_1 , interviewed by interviewer j_2 . The parameters associated with the independent variable x have a residual for both classifying structures. For this model the additional assumption is made that the residuals of the different classifying structures (in our case: the respondent and interviewer residuals) are mutually independent (u_{0j_1} and u_{1j_1} versus u_{0j_2} and u_{1j_2}).

Raudenbush (1993) discusses this kind of models and the use of the EM algorithm to estimate them. Rasbash and Goldstein (1994) and Goldstein (1995, 123-124) show how these models can be specified and estimated using a purely hierarchical formulation and (consequently) standard multi-level software. The way to do this is to specify one of the classifications as a standard hierarchical one, then define a dummy for each unit of the other classification, specify that each of these dummy variables has a random coefficient at the higher level and constrain the resulting sets of variances to be equal.

In section 6 and 7 we'll use these 3 different models. They can all be implemented in MLn/MLwiN, software for

multilevel modeling. Firstly we take a closer look at the variables we will use in the analysis.

5. VARIABLES IN OUR ANALYSIS

One of the more difficult tasks during the interview of the election study was rating six parties on different 11-point scales. Three scales were presented to the respondents: catholicism, economic liberalism and federalism. An explicit "don't know" filter was included in the question, but it was not mentioned on the card with the alternatives given to the respondent. The entire question is included in the Appendix. We expected a considerable number of "don't know" answers because of the degree of complexity of the task. The explicit filter was expected to raise that number as well (see *e.g.*, Schuman and Presser 1981).

In the first wave the average number turned out to run up to more than 4 "don't know" answers per respondent. Almost 20% of the respondents made use of this possibility at least 9 times out of the 18. If we consider only the panel respondents the mean number is a bit lower (3.8). This is not surprising since we could expect that "multi-users" of the "don't know" answer would be underrepresented in the second wave because of lack of interest in the subject of the survey and/or difficulties in answering the questions. In the second wave the overall mean is 3.6 and the mean for the panel respondents 3.4. The numbers for the respondents that were interviewed twice by the same interviewer are 3.9 and 4.2 respectively. There is no explanation why the number of "don't know" answers during the second wave is higher than the overall mean for these respondents.

At the measurement level we'll use the year of the interview as indication of time of measurement. We've recoded this variable, so time has the value 0 for the first wave and 3 for the second wave.

At the respondent level we have 3 independent variables: sex (0 = man, 1 = woman), completed education (0 = low, 1 = high) and the extent to which the respondents follow political news in the press (press: 1 = (almost) always - 5 never). The first 2 variables are constant for the 2 times of measurement. The third is a time-varying covariate and the question phrasing also slightly changed for the second survey. The two different questions are also included in the appendix. The dissimilarity in the phrasing induces an additional difficulty in setting up the model. The way to handle such a variable is to standardize it (mean 0, variance 1) for each time of measurement and (afterwards) to ascribe the value 0 to the time of measurement when the question wasn't asked. The reference value for those variables is their mean (see Snijders 1996, 422). This gives us 2 variables: press1 for the first occasion and press2 for the second. The former has the value 0 for all respondents for the second measurement and the latter for the first measurement. We don't take up the respondent's age in the model,

since this variable would correlate too much with the time measurement at the occasion level.

In order not to complicate the analysis too much we don't take up interviewer variables. We just assume there is an interviewer effect, without trying to explain that effect in terms of interviewer characteristics.

6. ANALYSIS 1: RESPONDENT TYPE 1 OF TABLE 1

The first analysis considers only those respondents who are interviewed twice by the same interviewer (cfr. Type 1 from Table 1). This analysis requires a "simple" three level model: measurements nested in respondents nested in interviewers. The hierarchical structure is unambiguous. This model is similar to the example in chapter 8 in the Bryk and Raudenbush book (1992). In that example the authors analyze the progress in academic achievement of students in schools.

Our dependent variable is the number of "don't knows" for respondent i on moment t , interviewed by j (Y_{tij}). We have only 2 measurements so the degree of the polynomial cannot exceed 1. This results in the following level 1 equation:

$$Y_{tij} = \pi_{0ij} + \pi_{1ij} \text{YEAR} + e_{tij}.$$

Our time variable (t) is the year of the interview which has the value 0 (1992) or 3 (1995). We will test whether π_{1ij} is significant. If not, this leaves us a null model or "naïve" model (see Snijders 1996, 411), in which the

number of "don't know" answers doesn't change over time, and both measurements can be considered as retests of the same constant value. The coefficients in the level 1 equation are respondent and interviewer specific.

At the respondent level we'll include 3 variables: sex, education and the 2 press variables. So our level 2 equation contains 4 variables:

$$\begin{aligned} \pi_{0ij} = & \pi_{0j} + \beta_{01j} \text{SEX}_i + \beta_{02j} \text{EDUCATION}_i \\ & + \beta_{03j} \text{PRESS1}_i + \beta_{04j} \text{PRESS2}_i + r_{0ij}. \end{aligned}$$

If the parameter estimate associated with year is significant we'll have a similar equation for π_{1ij} .

At the third level (interviewer) we won't include any more variables, but we will fit a random intercept and random slopes. So we have the following level 3 equations:

$$\pi_{0j} = \pi_0 + u_{0j} \text{ and } \beta_{01j} = \beta_{01} + u_{01j}, \dots$$

Implementing these model specifications in MLn gives us the following results.

Model a in the table is the null model. This is a model without independent variables, neither at the measurement level, nor at the respondent level. In this model there is no evolution in the number of "don't know" answers. But the variance of the dependent variable is divided in a measurement part, a respondent part and an interviewer part. All variances are significant. This indicates that there is between wave variation, that some respondents use the answer more than others and that some interviewers will get more "don't know" answers than other interviewers.

Table 2
Analysis of Respondents who were Interviewed Twice by the Same Interviewers (s.e. in brackets)

	model a	model b	model c	model d
Fixed				
Measurement level				
constant	4.136 (0.322)	4.028 (0.358)	3.749 (0.442)	3.754 (0.523)
year		0.072 (0.089)		
Respondent level				
sex			2.393 (0.434)	2.458 (0.414)
education			-1.675 (0.425)	-1.778 (0.446)
press1			0.911 (0.263)	0.887 (0.233)
press2			1.483 (0.236)	1.426 (0.234)
Random				
Interviewer level				
$\sigma^2_{\text{constant}}$	2.249 (1.040)	2.251 (1.043)	2.666 (0.969)	6.090 (2.109)
$\sigma^2_{\text{education/constant}}$				-4.099 (1.816)
$\sigma^2_{\text{education}}$				1.396 (1.819)
Respondent level				
$\sigma^2_{\text{constant}}$	14.470 (1.714)	14.480 (1.714)	8.939 (1.308)	8.692 (1.332)
Measurement level				
σ^2_e	13.320 (0.974)	13.300 (0.974)	13.270 (0.969)	13.250 (0.969)
-2 LL	4519.35	4518.62	4414.52	4395.68
Δdf^*		1	4	6

Note: * compared to model a

The inclusion of the variable YEAR does not provide a better fit of the model. The decrease in deviance ($-2 \log L$) is not significant, neither is the parameter of the variable significant (model b). We can conclude that there is no significant overall evolution in the number of "don't know"s. We can go on with a model without the time variable.

The respondent variables do result in a considerable improvement of fit of the model. The decrease of the $-2 \log L$ value is large and clearly significant ($p < 0.001$). According to the analysis women use the "don't know" alternative more than men do and highly educated respondents less than respondents with lower education. Following the political news in the press reduces your chance to answer "don't know". Both press1 and press2 are significant (model c). The inclusion of respondent variables also results in a substantial decrease of the variance at the respondent level.

We also tried to fit random slopes at the interviewer level (model d). Our analysis showed some variation in the parameter associated with the respondent's education. This is the only independent variable with a varying coefficient at the third level. $\sigma^2_{\text{education}}$ is not significant, but there is an important covariance between the residual for the constant and the residual for education ($\sigma_{\text{education/constant}} = -4.099$). The covariance is negative, indicating that interviewers with a higher constant have a smaller coefficient for education. Since the fixed parameter for education is negative, it will be even more negative for those interviewers, thus having a larger absolute value. Hence for interviewers who stimulate more "don't know" answers, the difference between less educated respondents and more educated respondents will be larger. In model d the value of $\sigma^2_{\text{constant}}$ at the interviewer level has increased considerably, compared to model c. In this model the variance at the interviewer level is dependent on the values of the explanatory variable education and it will be larger for zero values of education. That is another interpretation of model d: the variance between interviewers is much higher for lower educated respondents than for higher educated respondents. This model with a more complex variance structure at level 3 has a better fit than the previous models.

When including YEAR in model c or model d, it turned out to be not significant either. Also in our final models there is no evidence for an evolution in the number of "don't know"s between the two waves. All models prove a significant interviewer effect. But the relative size of the variance shows that there is more variation between respondents than between interviewers.

7. ANALYSIS 2: ALL RESPONDENTS

In this analysis we look at all the respondents: the panel respondents that were interviewed twice by the same interviewer, the other panel respondents and those who were

interviewed only once. This second analysis breaks down the hierarchical structure. Measurements are still nested in respondents and respondents are still nested in interviewers. But there is no overall hierarchical structure, since the interviewer can (and most of the time will) change between the two waves (see section 3). Our dependent variable is still the number of "don't know"s of respondent i interviewed by j on moment t (Y_{ijt}). But the model has changed. The level 1 equation hasn't:

$$Y_{ijt} = \pi_{0ij} + \pi_{1ij} \text{YEAR} + e_{ijt}.$$

In this notation we use π , since the level 1 model is also a growth curve. But this equation matches the level 1 model of the cross-classified model (equation (8), section 4.3). Furthermore we still use i for the respondent and j for the interviewer. But it is important to notice that this is not the same model as the one of analysis 1. These subscripts correspond to the j_1 and j_2 of equations (8) and (9).

There is no "real" third level. To fit the cross-classified model in MLn, we have to define a third level, but conceptually the respondent and interviewer are at the same level in this model. This leads to the following level 2 equation:

$$\begin{aligned} \pi_{0ij} = & \pi_0 + \beta_{01} \text{SEX}_i + \beta_{02} \text{EDUCATION}_i \\ & + \beta_{03} \text{PRESS1}_i + \beta_{04} \text{PRESS2}_i + r_{0i} + r_{0j}. \end{aligned}$$

The interviewer specific part (r_{0j}) is included in the second level, so there is no interaction between the interviewer variance and the respondent variables. This is the main difference with analysis 1.

A cross-classified model requires enormous computations. We have 3,026 respondents and 275 interviewers. This would mean 275 dummies with all varying coefficients at the artificial third level. Up till now it is impossible to fit such a model. The storage required by the worksheet is far too large (see Goldstein 1995, 118 and Rasbash and Woodhouse 1996, 85-86 for details). It is possible to reduce these storage requirements and improve the speed of model estimation by dividing the dataset in subsets in which the cross-classification implies fewer cells. In our case we look for separate groups of measurements that are classified by fewer respondents and interviewers. The analysis of 1 group of 1,000 measurements classified by 500 respondents and 100 interviewers is computationally more demanding than the analysis of a dataset consisting of 10 groups of 100 measurements each, classified by 50 respondents and 10 interviewers. Sometimes it is worth omitting some of the observations (measurements in combinations of respondents and interviewers that hardly occur) to make the partitioning more efficient.

MLn/MLwiN provides some procedures (via the commands XSEArch and BXSEArch) that are designed for that partition (Rasbash and Woodhouse 1996, 89-93). We used the BXSEArch command. The command starts an enhanced procedure, which attempts to provide the maximum separation with the minimum deletion of data. We started with

4,790 measurements, 3,026 respondents and 275 interviewers. After omitting the observations indicated by the BXSearch command, we're left with 4,597 measurements on 3,026 respondents interviewed by 275 interviewers. No higher level units (respondents nor interviewers) are left out. The procedure resulted in 7 partitions with a maximum of 44 cells in the cross-classification of respondents and interviewers. The model converged sufficiently fast when implied this way.

The results of the analysis are reported in Table 3.

Table 3
Analysis of all the Respondents (s.e. in brackets)

	model a	model b	model c
Fixed			
Measurement level			
constant	3.894 (0.136)	3.967 (0.155)	3.864 (0.165)
year		-0.053 (0.055)	
Respondent level			
sex			1.808 (0.153)
education			-1.914 (0.148)
press1			1.185 (0.090)
press2			1.197 (0.102)
Random			
Level 2			
Interviewer			
$\sigma^2_{\text{constant}}$	2.777 (0.373)	2.716 (0.368)	2.844 (0.363)
Respondent			
$\sigma^2_{\text{constant}}$	11.810 (0.635)	11.800 (0.635)	7.017 (0.527)
Measurement level			
σ^2_e	13.130 (0.475)	13.150 (0.476)	13.460 (0.480)
-2 LL	27717.1	27716.3	27042.1
Δdf^*			

Note: * compared to model a

This table looks very much the same as Table 2 but there is an important difference. In the random part we marked level 2 – interviewer and respondent to make clear that the interviewers do not constitute a third level in this analysis.

Model a is the null model: no explanatory variables, but the variance of the dependent variable separated in a measurement part, a respondent part and an interviewer part. There is a significant interviewer variance. Thus in this design we again have evidence for an interviewer effect. You have to be careful about the interpretation of the relative sizes of the variances if one classification has far fewer units than the other (Goldstein 1995, 117-118). It's not fully correct to state that there's 5 times as much variation between respondents than between interviewers, but again there is much more variability between respondents than between interviewers.

In the next model (model b) we've included the time variable (YEAR). Again this variable turns out to be not significant and its inclusion does not provide a better fit of the model. Again we can conclude that there is no significant overall evolution in "don't know" answers over time.

Model c is the model with the respondent variables. They are all significant and this model has a far better fit than the previous ones. The substantive interpretation of the parameters is the same as in analysis 1. Women use the "don't know" answer more than men and a higher education results in less "don't know"s. The extent to which the respondents follow the political news in the press is also a predictor of the use of the "don't know" answer. The less they follow politics the more they answer "don't know".

8. CONCLUSION AND DISCUSSION

The general conclusions of this article are methodological as well as substantive.

Our analysis confirms previous research findings about the use of the "don't know" answer. It is related to the respondent's education, sex and a measure of involvement or interest in the subject. Furthermore it is likely to diverge from interviewer to interviewer. All our analyses showed a significant interviewer effect. We did not find a significant evolution in the use of the "don't know" answer over time in the two waves of the survey. The interviewer effects prove that the "don't know" response alternative is not merely a result of the respondent answering the questions. It stresses the necessity of an interviewer training, which includes instructions on how to ask difficult questions and how to deal with "don't know" answers.

As in most panel surveys, the nonresponse in the second wave of this panel survey was not totally random. It is related to the respondent's living arrangement, his or her political interest and a few socio-demographic variables (Loosveldt and Carton 1997). This selective dropout puts limits to the generalizability of the results concerning the evolution in the dependent variable, but our analyses did not show a general evolution in the use of the "don't know" answer anyway. An impact of selective nonresponse in the second wave on the size of the interviewer effect is not unlikely either as interactions between the respondent characteristics and the interviewer effects are possible, as analysis 1 showed. But it is unlikely that this will affect the substantive conclusions about the interviewer effects. Given the results of analysis 1 and the conclusions in the Loosveldt and Carton paper, one could even expect that the interviewer effect in analysis 2 and consequently the overall interviewer effect might be somewhat underestimated. Loosveldt and Carton (1997, 1021) show that lower educated respondents are more likely to drop out of the survey than higher educated respondents and analysis 1 showed that the interviewer variance is higher for lower educated respondents.

The methodological conclusions consider the use of the different models to analyze interviewer effects in panel surveys. The analyses presented in this paper show that quite complex designs with complicated data structures can be analyzed by specifying the appropriate multilevel model.

The first model (Analysis 1) only suits in a tiny number of cases. It is not so common to ascribe the same interviewers to the same respondents for different waves of a panel survey, neither is it always feasible.

The second model (Analysis 2) is an appropriate tool but can require enormous computations. MLn is quite powerful and helps to decrease the storage requirements, at the cost of a small loss of information. Besides, the second model has its limitations too. Using this method it is not possible to model interactions between respondent variables and interviewer variance, as we did in the first analysis, or between respondent and interviewer variables. However the analysis showed that this model could be a very useful and flexible tool. The cross-classified model is also suitable when the number of measurements increases. A panel survey with 3 or 4 or even more waves, where some interviewers are retained and some are new at each occasion would require exactly the same analysis. The multilevel model also knows how to handle respondents for whom 1 or more measurements are missing, as our analysis showed. The pliability of this model outweighs the impossibility to include respondent – interviewer interactions in the model. That would be feasible when analyzing each wave of the panel survey separately. But those analyses could not model a possible evolution in the dependent variable, which is another important advantage of the joint analysis of all waves of the panel.

ACKNOWLEDGEMENTS

We thank the ISPO-PIOP Centre for Electoral Research for providing us with the data. Jacques Billiet, Marc Swyngedouw, Ann Carton and Roeland Beerten originally collected the Flemish data. The ISPO-PIOP is supported by the Federal Services for Technical, Cultural and Scientific Affairs. Neither the original collectors nor the Centre bear any responsibility for the analyses or interpretations presented here. We would also like to thank Jon Rasbash (Institute of Education, University of London) for some very useful comments about the operation of MLn. Finally we thank the referees for their constructive comments and suggestions on an earlier version of the paper.

APPENDIX 1

The rating question was: "Political parties are said to be "Catholic" or "non-Catholic". Please place the cards of the various parties on card No. 20 at the place that corresponds best to the degree in which the party is "Catholic" or "non-Catholic". If two or more parties are just as "Catholic" or just as "non-Catholic" in your opinion, place the cards on the same square. If you do not know how "Catholic" or "non-Catholic" a party might be, then simply put its card aside."

With the card:

Catholic 0 1 2 3 4 5 6 7 8 9 10 Non-Catholic

The press question was not identical for both surveys. For the first survey the press question was: "How often do you read the political news in the newspaper?"

With the response categories:

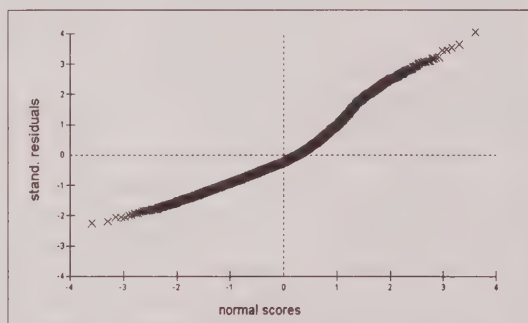
1=(almost) always, 2=often, 3=now and then, 4=seldom, 5=never

In the second survey it became: "How often do you follow the political news on the radio, on television or in the paper?"

The response categories remained the same.

APPENDIX 2

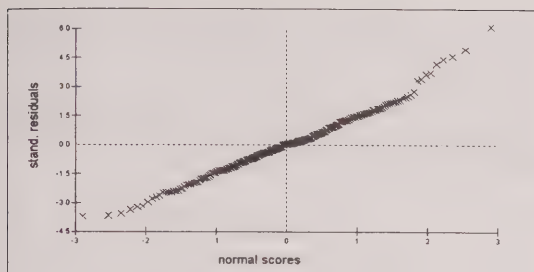
Section 4 set the assumptions of the different models that were used. For the last model the most important assumptions concern the random effects associated with the respondent and the interviewer. The assumption that the $\sigma^2_{\text{constant}}$ values for the respondent and for the interviewer are normally distributed can be assessed by looking at Normal probability plots for the residuals. Graph 1 presents the plot for the standardized respondent residuals and graph 2 the plot for the standardized interviewer residuals.



Graph 1. Standardized respondent residuals by Normal equivalent scores

In this graph the departures from the diagonal are rather limited and no apparent violation of Normality can be inferred. On the other hand it is worth noting that this graph shows more observations at the upper right hand than at the lower left end.

Graph 2 does not show any clear departures from the diagonal either. But in this graph some outliers draw the attention. Especially the outlier at the upper right hand side of the graph seems to be outside the range of the other interviewer residuals. Moreover in this graph also there are more observations at the upper right hand side than at the lower left end.



Graph 2. Standardized interviewer residuals by Normal equivalent scores.

The conclusions from these graphs are as follows: there is nothing clearly wrong with the residuals but the more numerous deviations upwards and the outliers of the interviewer residuals could possibly be further investigated. Efficient techniques for these checks are not yet available for multilevel models (Goldstein 1995: 29). But it is of course possible to analyze a dataset without the outliers. That is done in Table 4.

Table 4

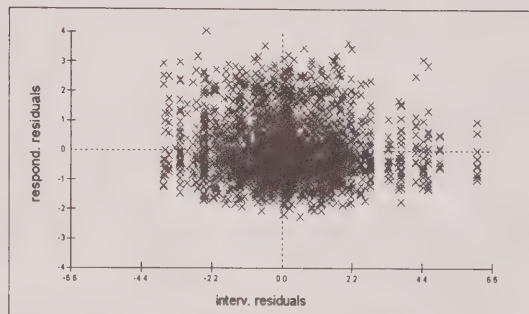
Analysis of the Dataset Without the Interviewer Outliers
(s.e. in brackets)

Fixed		
Measurement level		
constant	3.853	(0.162)
Respondent level		
sex	1.820	(0.153)
education	-1.929	(0.149)
press1	1.160	(0.090)
press2	1.217	(0.102)
Random		
Level 2		
Interviewer		
$\sigma^2_{\text{constant}}$	2.495	(0.333)
Respondent		
$\sigma^2_{\text{constant}}$	7.109	(0.530)
Measurement level		
σ^2_e	13.420	(0.481)
-2 LL	26850.2	

For the analysis in Table 4 we excluded two interviewers, the one with the lowest and the one with the highest residual. The coefficients in this table are very similar to those of model c in Table 3. The interviewer variance has decreased a bit, as a result of the exclusion of extremes, but there is no evidence of a considerable impact of the outliers on the results.

The other assumption about the interviewer and respondent random effects is their mutual independence. The interviewer and respondent residuals should not correlate. That is of course more difficult to evaluate since both residuals are connected to their respective units, which do not

correspond. You will get 3,028 respondent residuals and 275 interviewer residuals. An indirect check of this assumption is possible by attributing the interviewer residuals to the respondents. This is done in graph 3.



Graph 3. Standardized respondent residuals by standardized interviewer residuals

In this graph, again the more numerous deviations upwards and the interviewer outliers draw the attention. Apart from that, no pattern can be discerned. Because of the interviewer outliers, there are fewer observations at the right hand side of the graph. But the respondent residuals do not really tend to be smaller if the interviewer residuals are higher. Neither is there any evidence of the opposite.

The check in graph 3 is imperfect as it attributes the interviewer residuals to the respondents. A better alternative might be to fit a more complex model with an interaction term between the two random effects. Goldstein (1995, 119) proposes this model. A test for the model improvement due to the interaction term can give an indication for the presence of a correlation between the residuals. Another alternative is the insertion of an additional level (the region) above interviewers and respondents. That model would include a term for the regional variation, which could cause a correlation between the interviewer and respondent residuals. Snijders and Bosker (1999, 159-160) describe this model. But both models require a different parameterization with various sets of dummies. Their clarification calls for a paper in itself and is consequently outside the scope of this paper.

REFERENCES

- BAILAR, B., BAILEY, L., and STEVENS, J. (1977). Measures of interviewer bias and variance. *Journal of Marketing Research*, 14, 337-343.
- BEERTEN, R., BILLIET, J., CARTON, A., and SWYNGEDOUW, M. (1997). *1995 General Election Study Flanders-Belgium. Codebook and Questionnaire*. Leuven: ISPO/Departement Sociologie, K.U. Leuven.
- BRYK, A.S., and RAUDENBUSH, S. (1992). *Hierarchical Linear Models Applications and Data Analysis Methods*. Newbury Park - London: Sage.

- CARTON, A., SWYNGEDOUW, M., BILLIET, J., and BEERTEN, R. (1993). *Source Book of the Voters' Study in Connection with the 1991 General Election*. Leuven: Sociologisch Onderzoeksinstituut/ ISPO.
- DIPRETE, T.A., and FORRISTAL, J.D. (1994). Multilevel models: Methods and substance. *Annual Review of Sociology*, 20, 331-357.
- GOLDSTEIN, H. (1995). *Multilevel Statistical Models*. London: Edward Arnold.
- GROVES, R. M. (1989). *Survey Error and Survey Costs*. New York: Wiley.
- HANSON, R.H., and MARKS, E.S. (1958). Influence of the interviewer on the accuracy of survey results. *Journal of the American Statistical Association*, 53, 635-655.
- HOX, J.J. (1994). Hierarchical regression models for interviewer and respondent effects. *Sociological Methods and Research*, 22, 300-318.
- HOX, J.J., DE LEEUW, E.D., and KREFT, I.G. (1991). The effect of interviewer and respondent characteristics on the quality of survey data: a multilevel model. In *Measurement Errors in Surveys*. (Ed. P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, and S. Sudman). New York: Wiley, 439-461.
- ISPO/PIOP (1995). *1991 General Election Study Belgium. Codebook and Questionnaire*. Leuven: ISPO.
- KREFT, I.G., and DE LEEUW, J. (1998). *Introducing Multilevel Modeling*. London: Sage Publications.
- KROSNICK, J.A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5, 213-236.
- LOOSVELDT, G., and CARTON, A. (1997). Evaluation of nonresponse in the Belgian Election Panel Study '91 - '95. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1017-1022.
- RASBASH, J., and GOLDSTEIN, H. (1994). Efficient analysis of mixed hierarchical and cross-classified random structures using a multilevel model. *Journal of Educational Statistics*, 19, 337-350.
- RASBASH, J., and WOODHOUSE, G. (1996). *MLn Command Reference*. Version 1.0a. London: Multilevel Models Project. Institute of Education, University of London.
- RAUDENBUSH, S. W. (1993). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational Statistics*, 18, 321-349.
- SÄRNDAL, C., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- SCHUMAN, H., and PRESSER, S. (1981). *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording and Context*. New York: Academic Press.
- SCHWARZ, N., and SUDMAN, S. (1995). *Answering Questions*. San Francisco: Jossey-Bass.
- SNIJDERS, T. (1996). Analysis of longitudinal data using the hierarchical linear model. *Quality & Quantity*, 30, 405-426.
- SNIJDERS, T., and BOSKER, R. (1999). *Multilevel Analysis. An introduction to basic and advanced multilevel modeling*. London: Sage Publications.
- SUDMAN, S., and BRADBURN, N. (1974). *Response Effects in Surveys*. Chicago: Aldine.
- YANG, M., and GOLDSTEIN, H. (1996). Multilevel models for longitudinal data. In *Analysis of Change. Advanced Techniques in Panel Data Analysis*. (Ed. U. Engel, and J. Reinecke). Berlin - New York: Walter de Gruyter, 191-220.

Screen Design and Question Order in a CAI Instrument Results From a Usability Field Experiment

MAREK FUCHS¹

ABSTRACT

Screen design and questionnaire design affect the interviewer behavior in a CAI environment. Previous research has shown that interviewers can work more properly and efficiently if suitable functions and features are incorporated in the CAI instrument. Usability experiments with the household roster of two large government surveys have shown that using grids and tables is an important feature to facilitate the interviewer's performance. While these experiments were conducted under laboratory conditions, we have results from a first field experiment. In March of 1998 a CATI survey on immigrants was fielded in Germany (response rate 84%, $n = 501$). Four different versions of a household roster were compared in this production study, testing two different screen designs together with two different question orders in a 2x2 factor design. The four versions were randomly assigned to interviewers and respondents. Time measures were built into the CATI program, and 234 randomly selected interviews were video taped and analyzed according to a coding scheme. Based on the data we assessed the usability of different CAI design features. The results show that the screen design as well as the question order have a significant influence on interview duration and interviewer behaviors. Especially the grid based and topic based version allows the fastest performance in terms of time used to complete the instrument. Results from the coding data suggest that the differences between versions are due to specific interviewer and respondent behaviors. The data indicates that the grid based topic version enables a respondent oriented interviewer behavior, and thus allows the best interviewer performance in terms of duration.

KEY WORDS: Computer assisted interviewing; Usability Testing; Field experiment; Screen design; Question order.

1. INTRODUCTION

Computer assisted interviewing is on its way to becoming a standard survey technique (Couper, Baker, Bethlehem, Clark, Martin, Nicholls and O'Reilly 1998). In telephone surveys as well as with personal interviews, more and more studies are conducted using computer assisted interviewing techniques (CAI). Many of the large government surveys in the US are in the transition to CAI or have completed it already. Even in Europe, we observe a shift towards computer assisted interviewing (Schneid 1991; Fuchs 1994, 1995; Laurie and Moon 1997; Projektgruppe SOEP 1998) – even though, the methodological aspects of this development do not constitute the main focus of European research, so far.

Researchers and people responsible for fielding surveys rely on computer assisted interviewing for several reasons: (Sometimes it seems, however, that substantial arguments are less important than just a specific market rush towards CAI.)

- They hope to collect data of higher quality due to built-in consistency checks and range checks during the course of the interview.
- CAI provides the possibility to use automated skip patterns and allows to design more complex instruments without putting too much burden onto the interviewers.
- They hope to spend less time and money for interviewing and post-processing and decrease survey

budgets once the up-front investment for hardware and software is paid off.

- They hope to benefit from CAI's ability to read external data into the interview which is especially interesting with panel studies.

The general movement towards CAI is evaluated positively. Researchers and field directors benefit from it (Nicholls and deLeeuw 1996) and interviewers (Couper and Burt 1994) as well as respondents (Baker 1992), reveal a great deal of sympathy or at least acceptance. On the other hand, computer assisted interviewing has introduced some additional problems into the interview situation, too: in the early years methodological research was mainly concerned with hardware and software problems (see Couper, Groves and Kosary 1989; Weeks, 1992 for overviews). Instead, recent studies dealt with interview and respondent acceptance, interview duration, and usability issues (Couper *et al.* 1998 for an overview). The present paper contributes to this later discussion of "technology effects" (Fuchs, Couper and Hansen 2000).

2. THEORETICAL BACKGROUND

For the purpose of the following analysis the theoretical focus is mainly on two usability issues: (1) segmentation of the interview flow and (2) lack of interviewer flexibility.

¹ Dr. Marek Fuchs, Catholic University of Eichstätt, Department of Sociology, Ostenstrasse 26, 85071 Eichstätt, Germany. E-mail: marek.fuchs@ku-eichstaett.de.

1. Segmentation: in a CAPI environment the interviewer has an additional burden: the process of keying takes place in the interview situation. Usually, an interviewer reads a question, receives an answer, enters the data, presses [enter] and then the next screen with the following question appears. Compared to PAPI interviewers cannot look ahead and anticipate the next upcoming question while recording the answers to the previous one and they cannot start reading the next question before pressing [enter] – they cannot work simultaneously on both tasks. As a result of this procedure the interviewer respondent interaction is segmented by [enter] keys. So far we do not have quantitative evidence that this kind of segmentation harms the data or the interview situation. But it is argued that the interviewer loses the “big picture”, and the relevance of questions and their relationship to each other may be unclear (House 1985; Groves and Mathiowetz 1984).

Our findings from several series of usability tests in the lab concerning the screen layout of a household roster (Couper *et al.* 1997; Hansen, Couper and Fuchs 1998) led to the suggestion of a specific screen layout that allows the interviewer to develop a more complex understanding of the instrument, maintain the interaction with the respondent, and enter data at the same time: Two different versions of a series of questions were tested under laboratory conditions in terms of the time necessary in order to complete the questions and ease of use. We compared a so-called item based design with a grid based design. House and Nicholls (1988) distinguished between three approaches in screen design for computer assisted instruments: item based, screen based and form based design. In the item based approach one question and one input field are displayed at a time, and logic operations are performed in the transition from one item based screen to the next. This design is easy to program and focuses the interviewer's attention on the actual question. The screen based approach combines several items that need to be answered in sequence. All logic operations are executed after each item. On a form based screen, many items are presented at the same time in a table or grid and the interviewer may use the cursor keys to move from field to field and to complete them in any order.

The item version tested in our experiment matches the characteristics specified by House and Nicholls (1988) for a screen based approach. In contrast, the grid based design is best described as a form based instrument. It allows interviewers to record the information in the order chosen by the respondent, it provides the interviewer with a better overview of the instrument and it more easily allows updates and backups (for details see Couper *et al.* 1997). Also, the design matches the interviewers' demand for more questions on one screen – both for speed of administration and for context knowledge. The following graph gives an impression of an item based and a grid based CAI screen design.

We found evidence that the grid based design reduces the segmentation: interviewers could start reading the next upcoming question while still entering the data to the previous question. Even backing up seems to be easier within a grid design. On the other hand, we found only modest support for a grid based design in terms of time used to complete the task (for details see Couper *et al.* 1997). This leads to the question: what can we do to decrease segmentation and to further improve the efficiency of a household roster in terms of duration?

2. Lack of flexibility: The second feature that might cause problems in a computer assisted interview is the lack of flexibility. One of the advantages of a CAI instrument is the fact that an interviewer can hardly skip any questions. Although CAI instruments can make extensive use of skip patterns and filters, they apply a pre-defined question order. Usually, each question needs an [enter] key before the system goes on to the next screen. It is seen as an advantage that this rigid question order avoids any trouble the interviewer might have with the routing through the instrument, questions for specific respondents, filters and skip patterns and so on. He or she can abandon this task and focus on the administration of the actual items. On the other hand, this causes a very strict question order and provides the interviewer with little flexibility in terms of question order. A small example demonstrates this effect: most CAI instruments apply a question order to their household roster, where all items for one person are asked before the interviewer works through the same items for the next person (“person based design” see Couper *et al.* 1997; Fuchs 2001 or “grouped questions” see Moyer 1996). The CAI instrument, for example, might request the respondent's age, educational level, and other questions first before asking for the age of the respondent's wife. (This can be explained in part by the way computer programs and data bases work: households represent the main records and persons or other entities are treated as subrecords.) When completing the questions of a household roster it might happen (and in fact it happens quite often, see below) that the respondent provides not only the answer to the current question (e.g. “I'm 34 years old”) but also to a related question: “I'm 34 years old and my wife is 32 years old” or the respondent might answer “We are all Black” when asked about his or her own race (Oksenberg, Beebe, Blixt and Cannell 1992).

While working with a paper instrument it is an easy task for an interviewer to make immediate use of the additional information provided by the respondent. In case he or she answers, for instance, “We are all black” the interviewer can easily mark the appropriate check boxes for all household members at once. For someone interested in questionnaire design this leads to the following question: given the lack of flexibility in a computer assisted environment, what is the best question order for collecting information about all household members?

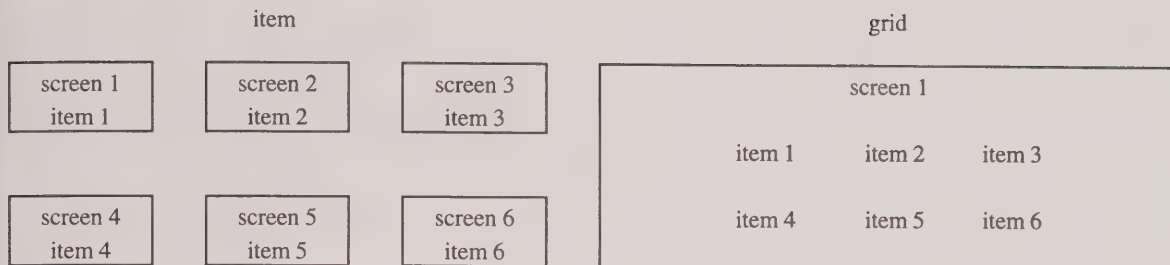


Figure 1. Item Based Design vs. Grid Based Design

Moore and Moyer report results from an experiment on two different question orders designed for collecting information about all eligible persons in a household (Moore and Moyer 1998a, 1998b). The first question order asks all questions for the first eligible person in the household and moves on to the next person, when all questions are completed. This question order is called a person based approach. In the second version, the topic based approach, the first question is asked for all eligible persons, then the second question for all persons and so on. Moore's and Moyer's results show strong support for a topic based design: the topic version leads to less item non-response, less break offs and refusals and is substantially shorter. Besides interviewers show significant preference for this version.

In the experiment presented in this paper we tried to make use of the advantages of a topic based approach and of a grid based screen design: we combined the two screen designs (item based design vs. grid based design) with the two question orders (person based order vs. topic based order) and tested all four resulting versions in a field experiment. In doing this, we had the following assumption in mind: the usability of a CAI instrument is not only a programming issue, but it is also connected to the questionnaire design and to the interview as a social situation. Both aspects of a computer assisted instrument, its screen design and its question order, support or hinder a smoothness of the interview flow. Based on the results of the previous research we had the following hypothesis: The combination of a grid based screen design and a topic based question order allows the most efficient interviewer respondent interaction.

3. METHODS

The experiment took place in Germany in March 1998. Immigrants of German origin from Poland, Rumania and the former Soviet Union were surveyed. Starting February 28, 1998 and ending March 20, 1998 15 interviewers completed $n = 501$ interviews. All respondents received an advanced letter and were called by phone up to 15 times.

The response rate reached 84% and item non-response was considerably low. The interviews were conducted using the CATI program CI3. About 95 questions on various topics were asked. The average interview lasted 23 minutes.

Four versions of a small household roster with three items per person were included in the instrument: an item/person version, a grid/person version, an item/topic version and a grid/topic version. All versions applied the same question wording and interviewer instructions, however, we modified the screen design and the question order according to the theoretical approach mentioned before (Figure 2). The item based person version is considered to be the standard version – it represents the questionnaire design usually applied to socio-demographic portions in CAI surveys. One of the four versions was randomly assigned to each interview – and thus to interviewers and respondents. We measured the total time needed for the household roster and in addition the time spent on each single item in that section of all 501 interviews. In addition, 234 interviews were selected at random and the interviewer working through the household roster section was videotaped. The video segments were coded in terms of interviewer behavior and respondent behavior and the resulting data was combined with the time measurements.

4. RESULTS

The durations of the four versions differ significantly from each other: interviewers needed 6.6 seconds per item in the item based person version (which is considered to be the standard one). In contrast each item took 5.5 seconds in the grid based topic version. This is a reduction of about 17% for the grid based topic version. The two other versions are in between.

It is important to mention that both factors seem to contribute to the decrease in time used to complete the task. If we distinguish between the two factors, we end up with the following results: the two topic based versions are significantly shorter than the two person based versions and the two grid based versions take significantly less time than the two person based versions. The combined effect applies to

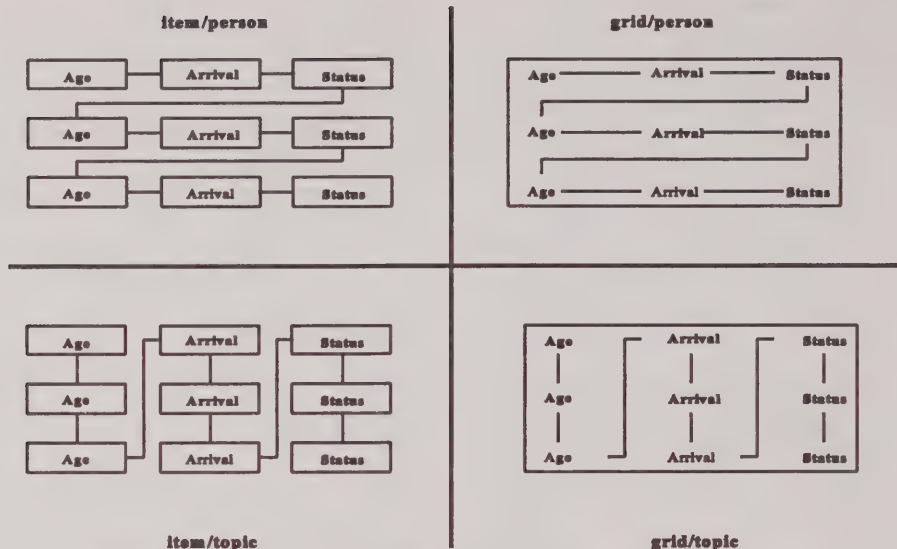


Figure 2. Four Versions Tested in the Experiment (Each Box Represents One Screen)

the grid based topic version and leads to the value of 5.5 seconds per item. (An analysis of variance reveals that both factors – the screen design as well as the question order – contribute independently to the decrease in time (screen design: $p < 0.01$, one third of total effect; question order: $p < 0.001$, two thirds of total effect, no significant interaction).)

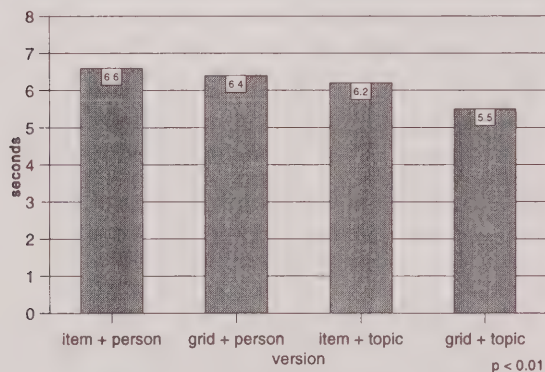


Figure 3. Duration per Item by Version

But why is the grid based topic version faster? A detailed analysis shows that this version is especially faster when collecting the information for the second and all following persons in the household – a significant impact, that is called loop effect (Fuchs 2001). This term describes the following phenomenon: the interviewer takes much longer to collect the information for the first person in a household compared to all subsequent persons. The average loop effect sums up to 3.4 seconds per item which is a reduction of about 38% compared to the first person (Table 1).

The loop effect is not specifically characteristic for this experiment. We recognized loop effects in our previous experiments with the NHIS household roster, too (Couper *et al.* 1997). It is, however, interesting to observe that the loop effect is significantly larger for the topic based versions than it is for the versions that follow a person based question order (Table 1). Thus the topic based versions do increase the acceleration for the second and all subsequent persons in a household and consequently show a larger loop effect. (One implication of our experimental design might be that interviewers did not know what version they were approaching. This may have decreased their performance on the very first item. But this effect should be the same across all versions, so the results should not be affected.)

Table 1
Duration and Loop Effect (Seconds)

Item	Age	Arrival	Status	All items
Duration per item				
First person in household	9.4	9.9	7.7	9.0
All other persons in household	6.6	5.7	4.5	5.6
All persons	8.0***	7.8***	6.1***	7.3***
Loop effect				
Differenz between first and all other persons in the household	-2.8	-4.2	-3.2	-3.4
Loop effect by version				
Grid + topic	-6.8	-6.4	-4.6	-5.9
Item + topic	-5.2	-8.3	-4.3	-6.0
Grid + person	-0.5	-2.7	-3.0	-2.1
Item + person	0.3	-0.3	-1.3	-0.4
Average loop effect	-2.8***	-4.2**	-3.2**	-3.4***

** $p < 0.01$; *** $p < 0.001$

Analyzing the video tapes we can provide reasons for these differences at least in part: given the topic based conditions, both interviewers and respondents adapt differently to the interview situation compared to the person based versions. When asking the questions for all persons in the household, the respondent recognizes the logic of the procedure very quickly. In quite a high proportion of all cases (about 30%) their reaction to this is "We all arrived in the same year" (meaning: "Don't ask me this question again and again").

If the instrument follows a person based design, the interviewer has to memorize this piece of information, and if it comes to the next person, he or she needs to remember: "Do not ask this question again, the respondent gave you the appropriate answer already!" Only in a few cases they really do, most of the time they just ask the question again. This is especially true when using an item based screen layout that gives no clues in terms of the answers to the same question for the other household members. In a topic based design instead, the interviewer can easily adapt to that situation. Thus he or she just enters the same code for all persons in the household without asking the question repeatedly. Both the interviewer and the respondent get used to the questions, and so the question answer process runs with less verbal contributions from the interviewer's side as well as from the respondent's. Both interviewer and respondent can anticipate the next question and the interview runs more smoothly. This is especially true when the CAI instrument makes use of a grid and provides further context information, *e.g.*, the responses for other household members to the same question. (Looking at the results reported in the lower part of Table 1 we conclude that the grid based person version does not benefit to the same extent from the advantages of the topic based approach. However, due to the grid design the loop-effect is considerably larger than in the item based person version.) As a result the time used per item is substantially shorter and the interviewer can provide respondent oriented interviewer behavior similar to Schober and Conrad's (1997) findings.

Providing feedback by the interviewer sometimes works as a signal that he or she has recorded the answer to the previous question in order to stimulate the respondent, so that the latter guesses about the next question and reveals the appropriate answer even without an additional stimulus. In extreme this might lead to a respondent behavior where he or she provides the information about all persons in the household at once: "We all came in the same year". The different versions tested in this experiment impel and support such behaviors to different degrees. From our results we can conclude that the grid based topic version stimulates interviewers and respondents to deviate from the scripted interview to a higher degree than the other versions. As far as duration is concerned this version allows the interviewer to make efficient use of information provided for all household members at once. Evidence from

the video coding support our interpretation of version-specific occurrences of time saving interviewer behaviors (1) and respondent behaviors (2):

1. By means of analyzing the video tapes we observe quite a lot of interviewer behaviors that do not follow standard interviewer procedures: besides the fact that about 78% of all items are read as worded, interviewers do not administer 9.3% of all items to the respondent. In another 5% of instances, the interviewer does not read the question but instead provides a different stimulus containing the relationship of the next person to the respondent (*e.g.*, "... and your wife?"). (It is interesting to recognize that interviewers chose the same verbal expressions on their own that Moore and Moyer 1998a, 1998b scripted in their experiments on question order.) In 5.5% of all cases the interviewer does not read the question but rather verifies the answer ("... and your wife is 32 years old?"). Some incomplete questions and wrong fills are observed, too. In total we have about 22% of all items affected by at least one interviewer behavior that does not follow a standardized interview script – which is a surprisingly high value considering that all interviewers were aware of the fact the interviews were video taped! Compared to other studies on interviewer behavior, however, the values are considerable lower. For example Oksenberg, Cannell and Blixt (1996) applied behavior coding to the National Medical Expenditure Survey and reported 37% to 41% of such interviewer behaviors. We will come back to the question of whether or not these behaviors help obtain valid measurements.

We draw the following conclusion from these particular findings: most of these behaviors indicate kind of a shortcut, *e.g.*, the interviewer does not read the question text as worded, he or she tries to make the conversation smoother and more suitable in terms of conversational rules. From our point of view this indicates that interviewers do not want to ask for information the respondent provided already. They do not want to behave unresponsively toward the verbal contributions of the respondent, instead, they wish to follow conversational rules. As a side effect these behaviors are less time consuming than standard interviewer behaviors. In our perspective, the priority therefore lies not with saving time, but with customizing the question answer process to respondent behaviors not anticipated and not absorbable by the computer assisted instrument.

In order to compare the four screen design versions in terms of the degree of interviewer deviations from the standard interview script we have computed the proportion of items per case affected by this kind of behavior. Large differences in interviewers not following the scripted interview between the four versions are to be noticed: Applying the grid based topic version to an interview results in more than twice as many such behaviors (the average proportion

of items affected is 0.48) than the item based person version (0.21) which is the standard for most studies so far. (An analysis of variance indicates that both factors contribute independently to the overall effect (screen design: $p < 0.001$; question order: $p < 0.001$; no significant interaction effect). About 25% of the overall effect can be attributed to the screen design, about three quarter to question order.) And this contributes to the time used for interviewing: items affected by a interviewer behavior not scripted in the interview take substantially less time (4.0 seconds) than the regularly administered items (6.8 seconds; $p < 0.001$).

Table 2
Interviewer Behavior and Respondent Behavior by Version

	Grid + topic	Item + topic	Grid + person	Item + person	Total
Average proportion of items affected by interviewer behavior not following the scripted interview per case	0.48	0.43	0.34	0.21	0.36***
Respondent provides information for all persons in the household at once	38.2%	44.4%	29.0%	10.8%	29.7%***

*** $p < 0.001$

In order to differentiate between the proportion of cases affected by a certain respondent behavior and the average proportion of items per case (!) affected by a certain interviewer behavior we used percent notation for the first and decimal notation for the later.

2. Additionally an analysis of the respondents' behavior shows that the topic design leads to a higher proportion of cases (42,3% compared to 19.7% for the person approach; $p < 0,001$) where the respondent provides at least once in the household roster section the information for all persons or a group of persons at once (e.g., "We all came in the same year"; "We all have the same legal status"). By contrast, the difference of the grid based design from the item based design is considerably smaller (33,6% vs. 26.1%) but does not reach the level of significance. However, an analysis of the interaction reveals a significant interaction effect ($p < 0.05$): Using a topic oriented question order the grid design does not make a significant difference. However, on top of an topic oriented question order the grid design increases the number of instances where the respondent provides the information for all household members at once.

It is surprising that results differ even for the two screen designs when using a person oriented question order. The study was administered by telephone, the respondents not being aware of the screen design at all. The only possible explanation is based on the fact that the interviewers modify their behavior in concordance with the screen design, stimulating the respondent differently. Accordingly, respondents, as well as interviewers, react to the screen design and the question order under the grid based person design in a way that facilitates the interviewer respondent interaction and thus helps smoothen the interview flow. (As seen before,

the interviewers change their behavior even under the grid based person condition (Table 2), however, the question order does not stimulate respondents to behave accordingly.)

One possible drawback of these interviewer and respondent behaviors might be a lack of data quality due to changes occurring in the predefined question answer process; instead, the respondent considers the answer less intensively and thoroughly. We observe only very few item missing values so an analysis of this standard indicator for data quality is not efficient. In fact, we do not expect a higher proportion of item missing values in either version. One might, however, be concerned about the homogeneity of the answers provided by the respondent. In a high proportion of cases he or she listens to the full question text only once and that could contribute to a less thorough consideration when answering the same question for subsequent household members. Additionally, answering for all household members at once ("We all arrived in the same year") might increase the homogeneity of the response and thus decrease data quality.

Table 3
Average Number of Different Categories (Homogeneity) per Household by Version

Variable	Grid + topic	Item + topic	Grid + person	Item + person	Total
Year of arrival (19 categories)	1.2	1.2	1.2	1.3	1.2
Status (4 categories)	1.3	1.3	1.3	1.3	1.3

No significant differences

In order to assess this possible drawback we computed the number of different response categories chosen by the respondent on a particular item for all household members (e.g., for year of arrival: respondent 1985, partner 1987, daughter 1987, son 1988 = 3 different response categories). This should give us an idea of whether or not only those respondent make use of the short-cut ("We came all in the same year") for whom this is actually valid, or whether even other respondents provided one answer for all household members even though they should have chosen two or more different response categories because of the situation in their particular household (unfortunately we have no external validation for the responses provided). In looking at the average number of different response categories (Table 3) we do not notice any differences in terms of homogeneity of data. For the year of arrival as well as for the legal status (as a German or a foreigner) there is no visible difference between the versions. For all versions the average number of different response categories chosen (one for each person in a household) shows no significant difference.

These finding provide only weak evidence that a grid design does not harm data quality. Other standard data quality indicators need to be assessed with larger data sets

in order to decide whether or not data quality is affected. However, based on the data available, we are unable to prove an effect on the validity of the responses.

5. DISCUSSION AND CONCLUSION

Our results from a comparison of four versions for a household roster (using the same question wording across versions) indicate that interviewers as well as respondents perform more efficiently under the grid based topic condition than with the other three versions. Combining a grid based screen design and a topic based question order reduces the average duration by about 17%. Two thirds of this reduction can be attributed to the question order, approximately one third to the screen layout. It is important to mention that the effect of the screen design is less pronounced than the one of question order and – compared to the effect on duration – even smaller on interviewer behavior and respondent behavior.

Even though the effects of the grid design on interviewer behavior and respondent behavior are far from large, they help to elicit two reasons for the better performance of the grid based topic version in terms of interview duration: (1) in the grid based topic version, the interviewer as well as the respondent adapt better to the logic of the question answer process, both anticipate the next question more easily and the question answer process runs more smoothly. (2) This version leads to more occurrences in which the respondent provides the information for the persons in the household faster and more often the respondents reveal the information for all household members or at least for one group at once. Even though the results are not fully consistent, this particular version makes it easier for the interviewer to adapt to this situation, record the information and stimulate the respondent to give the next appropriate answer without repeating the full question text.

Our findings contribute to the discussion of how to design survey instruments for interviewer administered computer assisted data collection. Based on the results reported in this paper we can draw the conclusion that making use of grids facilitates the interviewer respondent interaction and helps speed up data collection. Our experiments on item design vs. grid design conducted in the University of Michigan Survey Research Center's usability laboratory have shown that we can improve interviewer performance by providing grids (Couper *et al.* 1997). Moore and Moyer (1998a; 1998b) have demonstrated that one can improve interview efficiency by switching to a topic based question order, too. The present paper indicates that the interview situation benefit even more when combining both features.

Using grids and a topic based question order causes a greater amount of instances where the interviewer deviates from the scripted interview. From a rigid methodological point of view this might be seen as an important drawback,

especially, if the interviewer deviates from the standard interview script using global questions for all household members. For example, Martin (1999) showed a significant increase in the number of people enumerated in a household if extra questions were asked. In addition, Kindermann and colleagues (1997) demonstrated for non-household roster type questions, that additional cues on victimization significantly increase enumerations of crimes. Generalizing these results to global questions across persons, one would expect a decrease of data quality as interviewers use non-scripted behaviors that apply global questioning methods. However, the results reported in this article do not indicate that interviewers are using global questions and the author does not recommend to make extensive use of global questions when designing a survey instrument. Instead, the findings lead to a screen design that allows interviewers to make use of information reported when the respondent switches to a global mode and provides the information for several household members at a time. So, we do not want to encourage researcher to make extensive use of global questions and we do not want to see interviewer modify the scripted questions in order to ask global ones. However, when confronted with a respondent providing more information than actually asked for, the screen design of the CAI instrument should not prevent interviewers from making use of it.

A grid based design has been proven to facilitate the interviewers job with respect to this task, because it allows interviewers to adjust their behaviors in concordance with general conversational rules. Basic findings of behavior coding suggest that interviewers frequently deviate from specific interviewing procedures. "These changes often reflect adjustments made by the interviewers to meet the exigencies of the situation: to melt it more congenially with communications immediately preceding it, or to adjust to the respondent's particular situation" (Oksenberg *et al.* 1992: 3). This is especially necessary when respondents do not limit their answers to the information requested by the question, but elaborate it or provide additional information. "Avoiding the appearance of not paying attention to the respondent, interviewers in this situation frequently filled in the answer themselves without asking the question, or asking it only in part" (Oksenberg *et al.* 1992: 5). They thus try to switch to more respondent oriented procedures to avoid looking unresponsive. A grid based screen design and a topic oriented question order supports interviewers to interact according to these conversational rules and with respect to the interview situation's needs. This might be acceptable or even preferable as long as we are talking about factoid questions and as long as these interviewer behaviors do not harm data quality (e.g., leading question or probes).

What needs to be done in order to improve the computer assisted instrument in its supporting function for the interviewer respondent interaction: Our data suggests that the grid based topic version leads to a specific interview flow,

so that interviewer and respondent can easily adapt to it. Jeff Moore (1996; Moore and Moyer 1998a, 1998b) has shown that interviewers prefer the topic based version. By contrast, we know little about the respondents' satisfaction with that question order. Assessing their opinion about the different version is consequently an important goal. Moreover, we do not know whether this version matches the way in which information is stored in the respondents' brains. It might be, that respondents can easily adapt to this version, but that in terms of cognitive and social burden or in terms of correctness of answers it is not the right method. Additionally, we need to focus on the question whether or not we can transfer our findings from a household roster to other segments of a questionnaire. Right now we are conducting a series of field tests comparing different design solutions for factoid information other than household roster information and for attitude items. The versions tested in this experiment differ in the degree of contextual information provided to the interviewer while administering a particular item (previous questions, next questions *etc.*). The general question sounds: what happens if we use grids or form based screens more extensively? Under what conditions and circumstances does it help to improve interview efficiency and what are the limitations to this approach? However, it is too early to present any results at this time.

In addition, there are more unanswered questions that need to be addressed in future research. Personally I would like to suggest a specific approach to assess these questions assuming that computer assisted instrument design is of importance to different clients: researchers, interviewers and respondents. Of course, it is important that a CAI instrument meets the researcher's needs to obtain his or her measurements and also that the question answer process be well designed for each single item. However, in my view considering the social dimension of the interviewer respondent interaction and the behaviors in between single items is also a matter of importance. If the CAI instrument disturbs the social dimension of the measurement process it might harm even data quality. So far we do not know which approach allows the best compromise between validity and reliability of the measurement process on the one hand and a smooth short and non-embarrassing interview flow on the other hand. In order to find out to what respect a specific CAI screen design might harm data quality and how it helps save time, money and interviewer effort we need to conduct more usability studies.

To assess the questions mentioned above we do need more field experiments. Due to the fact that we want to analyze the social dimension of the interview and its effects on interviewer behavior as well as on interview duration, laboratory experiments do not meet our needs completely. Of course laboratory experiments allow a more controlled setting, reveal more detailed information about both participants, and – as a result – need smaller numbers of cases. Still, without going into the field, we will never confront

our prototypes and design solutions with real pressure to maintain and facilitate the interviewer respondent interaction and the question answer process at the same time. Usability testing should therefore be seen as a joint process of laboratory experiments and field tests.

ACKNOWLEDGEMENTS

Some of these results were presented at the SMP Brown Bag Seminar, Institute for Social Research, University of Michigan on May 21, 1998 and on the occasion of the 54th Annual Meetings of the American Association for Public Opinion Research, May 16, 1999. Special thanks go to the interviewers participating in this experiment. Mick Couper, Siegfried Lamnek, Jeffrey Moore and two anonymous reviewers provided helpful suggestions on earlier versions of this paper.

REFERENCES

- BAKER, R.P. (1992). New technology in survey research: computer-assisted personal interviewing (CAPI). *Social Science Computer Review*, 10, 145-157.
- COUPER, M.P., BAKER, R.P., BETHLEHEM, J., CLARK, C.Z.F., MARTIN, J., NICHOLLS, W.L., and O'REILLY, J. (Eds.) (1998). *Computer Assisted Survey Information Collection*. New York: Wiley.
- COUPER, M.P., and BURT, G. (1994). Interviewer attitudes toward computer-assisted personal interviewing (CAPI). *Social Science Computer Review*, 12, 38-54.
- COUPER, M.P., GROVES, R.M., and KOSARY, C. (1989). Methodological issues in CAPI. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 349-354.
- COUPER, M.P., FUCHS, M., HANSEN, S.E., and SPARKS, P. (1997). CAPI Instrument Design for the Consumer Expenditure (CE) Quarterly Interview Survey. Final Report. University of Michigan.
- FUCHS, M. (1994) *Umfrageforschung mit Telefon und Computer*. Einführung in die computergestützte telefonische Befragung. Weinheim: Psychologie Verlags Union.
- FUCHS, M. (1995). Die computergestützte telefonische Befragung. Einige Antworten auf Probleme der Umfrageforschung. *Zeitschrift für Soziologie*, 24, 284-299.
- FUCHS, M. (2001). The impact of technology on interaction in computer-assisted interviews. (Ed. D. W. Maynard, H. Houtkoop-Steenstra, N.C. Schaeffer, and H. van der Zouwen). *Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview*. Wiley (forthcoming).
- FUCHS, M., COUPER, M., and HANSEN, S. (2000). Technology effects: Do CAPI or PAPI interviews take longer? *Journal of Official Statistics* (in press).

- GROVES, R.M., and MATHIOWETZ, N.A. (1984). Computer assisted telephone interviewing: effect on interviewers and respondents. *Public Opinion Quarterly*, 48, 356-369.
- HANSEN, S.E., COUPER, M.P., and FUCHS, M. (1998). Usability Evaluation of the NHIS Instrument. Paper presented at the Annual Meeting of the AAPOR, St. Louis, MO.
- HOUSE, C.C. (1985). Questionnaire design with computer assisted telephone interviewing. *Journal of Official Statistics*, 1, 209-219.
- HOUSE, C.C., and NICHOLLS, W.L. (1988). Questionnaire design for cati: design objectives and methods. (Ed. R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls, and J. Waksberg). *Telephone Survey Methodology*, New York: Wiley, 421-436.
- LAURIE, H., and MOON, N. (1997). Converting to CAPI in a Longitudinal Panel Study. Working papers of the ESRC Research Centre on Micro-Social Change, 97-11, Essex.
- MARTIN, E. (1999). Who knows who lives here? Within-household disagreements as a source of survey coverage error. *Public Opinion Quarterly*, 63, 220-236.
- MOORE, J.C. (1996). Person- vs. Topic-based Design for Computer-Assisted Household Survey Instruments. Paper presented at InterCASIC '96, International Conference on Computer-Assisted Survey Information Collection, San Antonio, TX.
- MOORE, J.C., and MOYER, H.L. (1998a). ACS/CATI Person-Based/Topic-Based Field Experiment – Final Report. Center for Survey Methods Research, U.S. Bureau of the Census.
- MOORE, J.C., and MOYER, H.L. (1998b). Questionnaire Design Effects on Interview Outcomes. Paper presented at the Annual Meeting of the AAPOR, St. Louis, MO.
- MOYER, L.H. (1996). Which is better: grid listing or grouped questions design for data collection in establishment surveys? *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 986-990.
- NICHOLLS, W.L., and de LEEUW, E. (1996). Factors in acceptance of computer-assisted interviewing methods: a conceptual and historic review. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 758-763.
- OKSENBERG, L., BEEBE, T., BLIXT, S., and CANNELL, C. (1992). *Research on the Design and Conduct of the National Medical Expenditure Survey Interviews*. Final report. Survey Research Center, Ann Arbor, USA.
- OKSENBERG, L., CANNELL, C., and BLIXT, S. (1996). Analysis of Interviewer and Respondent Behavior in the Household Survey. U.S. Department of Health and Human Services. AHCPR No. 96-N016.
- PROJEKTGRUPPE SOEP (1998). Funktion und Design einer Ergänzungsstichprobe für das Sozio-ökonomische Panel. Diskussionspapiere des DIW, 163, Berlin.
- SCHNEID, M. (1991). Einsatz computergestützter Befragungssysteme in der Bundesrepublik Deutschland. Ergebnisse einer Umfrage. ZUMA-Arbeitsbericht 91/20. Mannheim: ZUMA.
- SCHÖBER, M.F., and CONRAD, F.G. (1997). Does conversational interviewing reduce survey measurement error? *Public Opinion Quarterly*, 61, 576-602.
- SUCHMAN, L., and JORDAN, B. (1990). Interactional troubles in face-to-face survey interviews. *Journal of the American Statistical Association*, 85, 45-54.
- WEEKS, M.F. (1992). Computer-assisted survey information collection: A review of CASIC methods and their implications for survey operations. *Journal of Official Statistics*, 8, 445-465.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following persons who have served as referees during 2000. An asterisk indicates that the person served more than once.

- J.-F. Beaumont, *Statistics Canada*
- W. Bell, *U.S. Bureau of the Census*
- * D.R. Bellhouse, *University of Western Ontario*
- Y. Berger, *University of Southampton*
- P. Biemer, *Research Triangle Institute*
- * D.A. Binder, *Statistics Canada*
- D. Cantor, *Westat Inc.*
- P. J. Cantwell, *U.S. Bureau of the Census*
- R.G. Carter, *Statistics Canada*
- J. Chen, *University of Waterloo*
- C.Z.F. Clark, *U.S. Bureau of the Census*
- M. Cohen, *National Center for Education Statistics*
- * J.-C. Deville, *Institut national de la statistique et des études économiques*
- * P. Dick, *Statistics Canada*
- * J.L. Eltinge, *U.S. Bureau of Labor Statistics*
- * W.A. Fuller, *Iowa State University*
- J. Gambino, *Statistics Canada*
- * M.A. Hidirolou, *Statistics Canada*
- D. Holt, *Office for National Statistics, U.K.*
- J.-S. Hwang, *Academia Sincia*
- * D. Judkins, *Westat, Inc.*
- * G. Kalton, *Westat, Inc.*
- S. Kaufman, *National Center for Education Statistics*
- J. Kim, *Westat Inc.*
- * P.S. Kott, *National Agricultural Statistics Service*
- * M. Kovačević, *Statistics Canada*
- M. Kramer, *U.S. Bureau of the Census*
- P. Lahiri, *University of Nebraska - Lincoln*
- N. Laniel, *Statistics Canada*
- * M. Latouche, *Statistics Canada*
- P. Lavallée, *Statistics Canada*
- S. Linacre, *Australian Bureau of Statistics*
- * S. Lohr, *Arizona State University*
- * H. Mantel, *Statistics Canada*
- P. Merkouris, *Statistics Canada*
- J. Moloney, *Statistics Canada*
- G. Nathan, *Central Bureau of Statistics, Israel*
- D. Norris, *Statistics Canada*
- J.-S. Pischke, *Massachusetts Institute of Technology*
- D. Pfeffermann, *Hebrew University*
- N. Plante, *L'Institut de la Statistique du Québec*
- N.G.N. Prasad, *University of Alberta*
- * B. Quenneville, *Statistics Canada*
- * E. Rancourt, *Statistics Canada*
- * J.N.K. Rao, *Carleton University*
- * L.-P. Rivest, *Université Laval*
- K. Rust, *Westat Inc.*
- I. Sande, *Telcordia Technologies*
- N. Schenker, *University of California - Los Angeles*
- F.J. Scheuren, *The Urban Institute*
- A. Scott, *University of Auckland*
- J. Sedransk, *Case Western Reserve University*
- * J. Shao, *University of Wisconsin - Madison*
- * M.P. Singh, *Statistics Canada*
- R. Sitter, *Simon Fraser University*
- C.J. Skinner, *University of Southampton*
- R.T. Smith, *National Agricultural Statistics Service*
- E. Stasny, *Ohio State University*
- * D. Stukel, *Statistics Canada*
- A. Théberge, *Statistics Canada*
- * Y. Tillé, *Ecole Nationale de la Statistique et de l'Analyse de l'Information*
- S. Tremblay, *Statistics Canada*
- C. Tucker, *U.S. Bureau of Labor Statistics*
- * R. Valliant, *Westat, Inc.*
- J. Waksberg, *Westat, Inc.*
- M. Wendt, *Statistics Canada*
- K.M. Wolter, *National Opinion Research Center*
- C. Wu, *University of Waterloo*
- * W. Yung, *Statistics Canada*
- * E. Zanutto, *University of Pennsylvania*
- A. Zaslavsky, *Harvard University*

Acknowledgements are also due to those who assisted during the production of the 2000 issues: J. Beauseigle (Dissemination Division) and L. Perreault (Official Languages and Translation Division). Finally we wish to acknowledge C. Corbeil, C. Ethier, C. Larabie, D. Lemire and G. Ray of Household Survey Methods Division, for their support with coordination, typing and copy editing.

Volume 28, No. 2, June/juin 2000, 225-448

James O. RAMSAY	
Differential equation models for statistical functions	225
Nancy E. HECKMAN and James O. RAMSAY	
Penalized regression with model-based penalties	241
Debbie J. DUPUIS and David C. HAMILTON	
Regression residuals and test statistics: assessing naive outlier deletion	259
Martin BILODEAU and Pierre DUCHESNE	
Robust estimation of the SUR model	277
Joris PINKSE	
Nonparametric two-step regression estimation when regressors and error are dependent	289
Ursula U. MULLER	
Nonparametric regression for threshold data	301
Ronald W. BUTLER and Aparna V. HUZURBAZAR	
Bayesian prediction of waiting times in stochastic models	311
Caterina CONIGLIANI, J. Ivan CASTRO and Anthony O'HAGAN	
Bayesian assessment of goodness of fit against nonparametric alternatives	327
Caterina CONIGLIANI and Anthony O'HAGAN	
Sensitivity of the fractional Bayes factor to prior distributions	343
William J. REED	
Reconstructing the history of forest fire frequency: identifying hazard rate change points using the Bayes information criterion	353
Antonio CUEVAS, Manuel FEBRERO and Ricardo FRAIMAN	
Estimating the number of clusters	367
Peter T. KIM and Ja-Yong KOO	
Directional mixture models and optimal estimation of the mixing density	383
Dominique FOURDRINIER and Idir OUASSOU	
Spherically symmetric distribution with constraints on the norm	399
Serge B. PROVOST and Young-Ho CHEONG	
On the distribution of linear combinations of the components of a Dirichlet random vector	417
Michael D. deB. EDWARDES	
Implications of random cut-points theory for the Mann-Whitney and binomial tests	427
Martin R. PETERSEN and James A. DEDDENS	
Effects of omitting a covariate in Poisson models when the data are balanced	439
Konstantinos FOKIANOS, Amy PENG and Jing QIN	
A generalized-moments specification test for the logistic link ²	446
Forthcoming Papers/Articles à paraître	447

Volume 28, No. 3, June/juin 2000, 449-672

Feifang HU & John D. KALBFLEISCH: The estimating function bootstrap	449
<i>Discussion:</i>	
James V. ZIDEK & Steven X. WANG: Comment 1	482
Thomas J. DICICCIO & Robert J. TIBSHIRANI: Comment 2	485
Christian LÉGER: Comment 3	487
Angelo J. CANTY & Anthony C. DAVISON: Comment 4	489
Stephen M. S. LEE: Comment 5	494
<i>Rejoinder:</i>	
Feifang HU & John D. KALBFLEISCH	496
Alan M. POLANSKY: Stabilizing bootstrap- <i>t</i> confidence intervals for small samples	501
Steven E. STERN & A. H. WELSH: Likelihood inference for small variance components	517
Francesca DOMINICI, Giovanni PARMIGIANI & Merlise CLYDE: Conjugate analysis of multivariate normal data with incomplete observations	533
Barbara TONG & Kert VIELE: Smooth estimates of normal mixtures	551
Dianliang DENG & Sudhir R. PAUL: Score tests for zero inflation in generalized linear models	563
Jiti GAO & Thomas YEE: Adaptive estimation in partially linear autoregressive models	571
Thomas S. FERGUSON, Christian GENEST & Marc HALLIN: Kendall's tau for serial dependence	587
Min CHEN & Gemai CHEN: Geometric ergodicity of nonlinear autoregressive models with changing conditional variances	605
Jean-Yves DAUXOIS: Inférence par les martingales pour des processus ponctuels à compensateur discontinu	615
Bing LI: Nonparametric estimating equations based on a penalized information criterion	621
Peter Xue-Kun SONG: Monte Carlo Kalman filter and smoothing for multivariate discrete state space models	641
Cheikh A. T. DIACK: Sur la convergence des tests de Schlee et de Yatchew	653
Forthcoming Papers/Articles à paraître	671
Volume 29 (2001): Subscription rates/Frais d'abonnement	672

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents

Volume 16, Number 1, 2000

Correcting the Bias in the Range of a Statistic Across Small Areas <i>David R. Judkins and Jun Liu</i>	1
A Procedure for Stratification by an Extended Ekman Rule <i>Dan Hedlin</i>	15
A Note on Use of Inverse Sampling: Post Estimation Between Successive Infections <i>Kung-Jong Lui</i>	31
Optimal Weighting of Index Components: An Application to the Employment Cost Index <i>Michael K. Lettau and Mark A. Loewenstein</i>	39
Random Selection in a National Telephone Survey: A Comparison of the Kish, Next-Birthday, and Last-Birthday Methods <i>Diane Binson, Jesse A. Canchola, and Joseph A. Catania</i>	53
The Effect of Different Rotation Patterns on the Revisions of Trend Estimates <i>David G. Steel and Craig H. McLaren</i>	61
Book and Software Reviews	77
In Other Journals	83

Volume 16, Number 2, 2000

Recent Developments for Poverty Measurement in U.S. Official Statistics <i>David M. Betson, Constance F. Citro, and Robert T. Michael</i>	87
Nearest Neighbor Imputation for Survey Data <i>Jiahua Chen and Jun Shao</i>	113
A Note on Jackknife Variance Estimation for the General Regression Estimator <i>Pierre Duchesne</i>	133
Stratification by Size Revisited <i>Alan H. Dorfman and Richard Valliant</i>	139
An Estimation File that Incorporates Auxiliary Information <i>Cary T. Isaki, M.M. Ikeda, J.H. Tsay, and Wayne A. Fuller</i>	155
Large Scale Fitting of Regression Models with ARIMA Errors <i>Björn Fischer and Christophe Planas</i>	173
Book and Software Reviews	185

Model-Based Alternatives to Trimming Survey Weights <i>Michael R. Elliott and Roderick J.A. Little</i>	191
Permanent and Collocated Random Number Sampling and the Coverage of Births and Deaths <i>Lawrence R. Ernst, Richard Valliant, and Robert J. Casady</i>	211
Survey Estimation for Highly Skewed Populations in the Presence of Zeroes <i>Forough Karlberg</i>	229
The General Application of Significance Editing <i>David Lawrence and Richard McKenzie</i>	243
Developing Usability Guidelines for AudioCasi Respondents with Limited Literacy Skills <i>Sid J. Schneider and Brad Edwards</i>	255
Technology Effects: Why Do CAPI Interviews Take Longer? <i>Marek Fuchs, Mick Couper, and Sue Ellen Hansen</i>	273
Book and Software Reviews	287
Editorial Collaborators	291

All inquires about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Department, Statistics Sweden, Box 24 300, S - 104 51 Stockholm, Sweden.

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue of *Survey Methodology* (Vol. 19, No. 1 and onward) as a guide and note particularly the points below. Accepted articles must be submitted in machine-readable form, preferably in WordPerfect. Other word processors are acceptable, but these also require paper copies for formulas and figures.

1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size ($8\frac{1}{2} \times 11$ inch), one side only, entirely double spaced with margins of at least $1\frac{1}{2}$ inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(-)" and "log(-)", etc.
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (e.g., w, ω ; o, O; l, 1).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 19, n° 1) et de noter les points ci-dessous. Les articles acceptés doivent être soumis sous forme de fichiers de traitement de texte, préférablement WordPerfect. Les autres logiciels sont acceptables, mais une version sur papier sera alors exigée pour le traitement des formules et des figures.

1. Présentation

- 1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.
- 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
- 1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
- 1.4 Les remerciements doivent paraître à la fin du texte.
- 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. Résumé

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. Rédaction

- 3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
- 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(-) et log(-) etc.
- 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
- 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
- 3.5 Distinguer clairement les caractères ambigus (comme w, ω; o, O, 0, I, 1).
- 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.

4. Figures et tableaux

- 4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
- 4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois).

5. Bibliographie

- 5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
- 5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

Model-Based Alternatives to Trimming Survey Weights <i>Michael R. Elliott and Roderick J.A. Little</i>	191
Permanent and Collocated Random Number Sampling and the Coverage of Births and Deaths <i>Lawrence R. Eyrns, Richard Valliant, and Robert J. Casady</i>	211
Survey Estimation for Highly Skewed Populations in the Presence of Zeros <i>Forough Karberg</i>	229
The General Application of Significance Editing <i>David Lawrence and Richard McKenzie</i>	243
Developing Usability Guidelines for AudioCasi Respondents with Limited Literacy Skills <i>Sid J. Schneider and Brad Edwards</i>	255
Technology Effects: Why Do CAPI Interviews Take Longer? <i>Marek Fuchs, Mick Couper, and Sue Ellen Hansen</i>	273
Book and Software Reviews	287
Editorial Collaborators	291

All inquiries about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Department, Statistics Sweden, Box 24 300, S - 104 51 Stockholm, Sweden.

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents

Volume 16, Number 1, 2000

1	David R. Judkins and Jun Liu	Correcting the Bias in the Range of a Statistic Across Small Areas
---	--	--

15	Dan Hedlin	A Procedure for Stratification by an Extended Ekman Rule
----	----------------------	--

31	Kung-Jong Lui	A Note on Use of Inverse Sampling: Post Estimation Between Successive Infections
----	-------------------------	--

39	Michael K. Lettau and Mark A. Loewenstein	Optimal Weighting of Index Components: An Application to the Employment Cost Index
----	---	--

53	Diane Binson, Jesse A. Canchola, and Joseph A. Catania	Random Selection in a National Telephone Survey: A Comparison of the Kish, Next-Birthday, and Last-Birthday Methods
----	--	---

61	David G. Steel and Craig H. McLaren	The Effect of Different Rotation Patterns on the Revisions of Trend Estimates
----	---	---

77		Book and Software Reviews
----	--	---------------------------

83		In Other Journals
----	--	-------------------

Volume 16, Number 2, 2000

87	David M. Belson, Constance F. Citro, and Robert T. Michael	Recent Developments for Poverty Measurement in U.S. Official Statistics
----	--	---

113	Jiahua Chen and Jun Shao	Nearest Neighbor Imputation for Survey Data
-----	------------------------------------	---

133	Pierre Duchesne	A Note on Jackknife Variance Estimation for the General Regression Estimator
-----	---------------------------	--

139	Alan H. Dorfman and Richard Valliant	Stratification by Size Revisited
-----	--	----------------------------------

155	Cary T. Isaki, M. M. Ikeda, J. H. Tsay, and Wayne A. Fuller	An Estimation Rule that Incorporates Auxiliary Information
-----	---	--

173	Björn Fischer and Christophe Planas	Large Scale Fitting of Regression Models with ARIMA Errors
-----	---	--

185		Book and Software Reviews
-----	--	---------------------------

Volume 28, No. 3, June/juin 2000, 449-672

449	Feifang HU & John D. KALBFLEISCH: The estimating function bootstrap
	<i>Discussion:</i>
482	James V. ZIDEK & Steven X. WANG: Comment 1
485	Thomas J. DICICCO & Robert J. TIBSHIRANI: Comment 2
487	Christian LÉGER: Comment 3
489	Angelo J. CANTY & Anthony C. DAVISON: Comment 4
494	Stephen M. S. LEE: Comment 5
	<i>Rejoinder:</i>
496	Feifang HU & John D. KALBFLEISCH
501	Alan M. POLANSKY: Stabilizing bootstrap- t confidence intervals for small samples
517	Steven E. STERN & A. H. WELSH: Likelihood inference for small variance components
533	Francesca DOMINICI, Giovanni PARMIGIANI & Muriel CLYDE: Conjugate analysis of multivariate normal data with incomplete observations
551	Barbara TONG & Kert VIELE: Smooth estimates of normal mixtures
563	Dianliang DENG & Sudhir R. PAUL: Score tests for zero inflation in generalized linear models
571	Jiti GAO & Thomas YEE: Adaptive estimation in partially linear autoregressive models
587	Thomas S. FERGUSON, Christian GENEST & Marc HALLIN: Kendall's tau for serial dependence
605	Min CHEN & Gemai CHEN: Geometric ergodicity of nonlinear autoregressive models with changing conditional variances
615	Jean-Yves DAVUXOIS: Inférence par les martingales pour des processus ponctuels à compensateur discontinu
621	Bing LI: Nonparametric estimating equations based on a penalized information criterion
641	Peter Xue-Kun SONG, Monte Carlo Kalman filter and smoothing for multivariate discrete state space models
653	Cheikh A. T. DIACK: Sur la convergence des tests de Schlee et de Yatcew
671	Forthcoming Papers/Articles à paraître
672	Volume 29 (2001): Subscription rates/Frais d'abonnement

James O. RAMSAY	Differential equation models for statistical functions	225
Nancy E. HECKMAN and James O. RAMSAY	Penalized regression with model-based penalties	241
Debbie J. DUPUIS and David C. HAMILTON	Regression residuals and test statistics: assessing naive outlier deletion	259
Martin BILLODEAU and Pierre DUCHESNE	Robust estimation of the SUR model	277
Jonis PINKSE	Nonparametric two-step regression estimation when regressors and error are dependent	289
Ursula U. MÜLLER	Nonparametric regression for threshold data	301
Ronald W. BUTLER and Aparna V. HUZURBAZAR	Bayesian prediction of waiting times in stochastic models	311
Caterina CONIGLIANI, J. Ivan CASTRO and Anthony O'HAGAN	Bayesian assessment of goodness of fit against nonparametric alternatives	327
Caterina CONIGLIANI and Anthony O'HAGAN	Sensitivity of the fractional Bayes factor to prior distributions	343
William J. REED	Reconstructing the history of forest fire frequency: identifying hazard rate change points using the Bayes information criterion	353
Antonio CUEVAS, Manuel FERRERO and Ricardo FRAIMAN	Estimating the number of clusters	367
Peter T. KIM and Ja-Yong KOO	Directional mixture models and optimal estimation of the mixing density	383
Dominique FOURDRINIER and Idir OUASSOU	Spherically symmetric distribution with constraints on the norm	399
Serge B. PROVOST and Young-Ho CHBONG	On the distribution of linear combinations of the components of a Dirichlet random vector	417
Michael D. deB. EDWARDS	Implications of random cut-points theory for the Mann-Whitney and binomial tests	427
Martin R. PETERSEN and James A. DEDDEN	Effects of omitting a covariate in Poisson models when the data are balanced	439
Konstantinos FOKIANOS, Amy PENG and Jing QIN	A generalized-moments specification test for the logistic link ²	446
	Forthcoming Papers/Articles à paraître	447

REMERCIEMENTS

Techniques d'enquête désire remercier les personnes suivantes, qui ont accepté de faire la critique d'un article durant l'année 2000. Un astérisque indique que la personne a participé plus d'une fois.

- J.-F. Beaumont, *Statistique Canada*
W. Bell, *U.S. Bureau of the Census*
D.R. Bellhouse, *University of Western Ontario*
Y. Berger, *University of Southampton*
P. Biemer, *Research Triangle Institute*
D.A. Binder, *Statistique Canada*
D. Cantor, *Westat Inc.*
P. J. Cantwell, *U.S. Bureau of the Census*
R.G. Carter, *Statistique Canada*
J. Chen, *University of Waterloo*
C.Z.F. Clark, *U.S. Bureau of the Census*
M. Cohen, *National Center for Education Statistics*
J.-C. Deville, *Institut national de la statistique et des études économiques*
P. Dick, *Statistique Canada*
J.L. Ettinge, *U.S. Bureau of Labor Statistics*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistique Canada*
M.A. Hidiroglou, *Statistique Canada*
D. Holt, *Office for National Statistics, U.K.*
J.-S. Hwang, *Academia Sinica*
D. Judkins, *Westat, Inc.*
G. Kalton, *Westat, Inc.*
S. Kauffman, *National Center for Education Statistics*
J. Kim, *Westat Inc.*
P.S. Kott, *National Agricultural Statistics Service*
M. Kovachevic, *Statistique Canada*
M. Kramer, *U.S. Bureau of the Census*
P. Lahiri, *University of Nebraska - Lincoln*
N. Lanier, *Statistique Canada*
M. Latouche, *Statistique Canada*
P. Lavallée, *Statistique Canada*
S. Linacre, *Australian Bureau of Statistics*
S. Lohr, *Arizona State University*
H. Mantel, *Statistique Canada*
P. Merkouris, *Statistique Canada*
J. Moloney, *Statistique Canada*
- G. Nathan, *Central Bureau of Statistics, Israel*
D. Norris, *Statistique Canada*
J.-S. Pischke, *Massachusetts Institute of Technology*
D. Pfeffermann, *Hebrew University*
N. Planie, *L'Institut de la Statistique du Québec*
N.G.N. Prasad, *University of Alberta*
B. Quenneville, *Statistique Canada*
E. Rancourt, *Statistique Canada*
J.N.K. Rao, *Carleton University*
L.-P. Rivest, *Université Laval*
K. Rust, *Westat Inc.*
I. Sande, *Telcordia Technologies*
N. Schenker, *University of California - Los Angeles*
F.J. Scheuren, *The Urban Institute*
A. Scott, *University of Auckland*
J. Sedransk, *Case Western Reserve University*
J. Shao, *University of Wisconsin - Madison*
M.P. Singh, *Statistique Canada*
R. Sitter, *Simon Fraser University*
C.J. Skinner, *University of Southampton*
R.T. Smith, *National Agricultural Statistics Service*
E. Stasny, *Ohio State University*
D. Sukel, *Statistique Canada*
A. Théberge, *Statistique Canada*
Y. Tillé, *Ecole Nationale de la Statistique et de l'Analyse de l'Information*
S. Tremblay, *Statistique Canada*
C. Tucker, *U.S. Bureau of Labor Statistics*
R. Valliant, *Westat, Inc.*
J. Wakseberg, *Westat, Inc.*
M. Wendi, *Statistique Canada*
K.M. Wolter, *National Opinion Research Center*
C. Wu, *University of Waterloo*
W. Yung, *Statistique Canada*
E. Zanutto, *University of Pennsylvania*
A. Zaslavsky, *Harvard University*

On remercie également ceux qui ont contribué à la production des numéros de la revue pour 2000: J. Beauséjour (Division de la diffusion) et L. Perteault (Division des langues officielles et traduction). Finalement on désire exprimer notre reconnaissance à C. Corbeil, C. Ehiher, C. Larabie, D. Lemire et G. Ray de la Division des méthodes des enquêtes auprès des ménages, pour leur apport à la coordination, la dactylographie et la rédaction.

HANSEN, S.E., COOPER, M.P., et FUCHS, M. (1998). Usability Evaluation of the NHIS Instrument. Article présenté à l' Annual Meeting of the AAPOR, St. Louis, MO.

HOUSE, C.C. (1985). Questionnaire design with computer assisted telephone interviewing. *Journal of Official Statistics*, 1, 209-219.

HOUSE, C.C., et NICHOLLS, W.L. (1988). Questionnaire design for cat: design objectives and methods. (R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls, et J. Waksberg, Eds.). *Telephone Survey Methodology*. New York: Wiley, 421-436.

LAVRIE, H., et MOON, N. (1997). Converting to CAPI in a Longitudinal Panel Study. Document de travail de l' ESRC Research Centre on Micro-Social Change, 97-11, Essex.

MARTIN, E. (1999). Who knows who lives here? Within-household disagreements as a source of survey coverage error. *Public Opinion Quarterly*, 63, 220-236.

MOORE, J.C. (1996). Person- vs. Topic-based Design for Computer-Assisted Household Survey Instruments. Article présenté à la InterCASIC '96, International Conference on Computer-Assisted Survey Information Collection, San Antonio, TX.

MOORE, J.C., et MOYER, H.L. (1998a). ACS/CAT Person-Based/Topic-Based Field Experiment – Final Report. Center for Survey Methods Research, U.S. Bureau of the Census.

MOORE, J.C., et MOYER, H.L. (1998b). Questionnaire Design Effects on Interview Outcomes. Article présenté à l' Annual Meeting of the AAPOR, St. Louis, MO.

MOYER, L.H. (1996). Which is better: grid listing or grouped questions design for data collection in establishment surveys? *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 986-990.

NICHOLLS, W.L., et de LEEUW, E. (1996). Factors in acceptance of computer-assisted interviewing methods: a conceptual and historic review. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 758-763.

OKSENBERG, L., BEEBE, T., BLIXT, S., et CANNELL, C. (1992). *Research on the Design and Conduct of the National Medical Expenditure Survey Interviews*. Rapport final. Survey Research Center, Ann Arbor, USA.

OKSENBERG, L., CANNELL, C., et BLIXT, S. (1996). Analysis of Interviewer and Respondent Behavior in the Household Survey. U.S. Department of Health and Human Services. AHCPR No. 96-N016.

PROJEKTGRUPPE SOEP (1998). Funktion und Design einer Ergänzungssstichprobe für das Sozio-oekonomische Panel. Diskussionspapiere des DIW, 163, Berlin.

SCHNEID, M. (1991). Einsatz computergestützter Befragungssysteme in der Bundesrepublik Deutschland. Ergebnisse einer Umfrage. ZUMA-Arbeitsbericht 91/20. Mannheim: ZUMA.

SCHOBER, M.F., et CONRAD, F.G. (1997). Does conversational interviewing reduce survey measurement error? *Public Opinion Quarterly*, 61, 576-602.

SUCHMAN, L., et JORDAN, B. (1990). Interactional troubles in face-to-face survey interviews. *Journal of the American Statistical Association*, 85, 45-54.

WEEKS, M.F. (1992). Computer-assisted survey information collection: A review of CASIC methods and their implications for survey operations. *Journal of Official Statistics*, 8, 445-465.

facilité d'utilisation à l'aide d'une combinaison d'expériences de laboratoire et d'essais sur le terrain.

REMERCIEMENTS

Une partie des résultats a été présentée à un colloque tenu à l'Institute for Social Research de l'Université du Michigan le 21 mai 1998, et à l'occasion de la 54^e assemblée annuelle de l'American Association for Public Opinion Research le 16 mai 1999. Nous remercions les intervieweurs qui ont participé à la présente expérience. Mick Couper, Siegfried Lamnek, Jeffrey Moore et deux examinateurs anonymes ont fourni des suggestions utiles au sujet de versions antérieures du présent exposé.

BIBLIOGRAPHIE

- BAKER, R.P. (1992). New technology in survey research: computer-assisted personal interviewing (CAPI). *Social Science Computer Review*, 10, 145-157.
- COUPER, M.P., BAKER, R.P., BETHLEHEM, J., CLARK, C.Z.F., MARTIN, J., NICHOLLS, W.T., et O'REILLY, J. (Eds.) (1998). *Computer Assisted Survey Information Collection*. New York: Wiley.
- COUPER, M.P., et BURR, G. (1994). Interviewer attitudes toward computer-assisted personal interviewing (CAPI). *Social Science Computer Review*, 12, 38-54.
- COUPER, M.P., GROVES, R.M., et KOSARY, C. (1989). *Methodological issues in CAPI. Proceedings of the Section on Survey Research Methods, American Statistical Association*, 349-354.
- COUPER, M.P., FUCHS, M., HANSEN, S.E., et SPARKS, P. (1997). CAPI Instrument Design for the Consumer Expenditure (CE) Quarterly Interview Survey. Rapport final. University of Michigan.
- FUCHS, M. (1994) *Umfrageforschung mit Telefon und Computer*. Einführung in die computergestützte telefonische Befragung. Weinheim: Psychologie Verlags Union.
- FUCHS, M. (1995). Die computergestützte telefonische Befragung. Einige Antworten auf Probleme der Umfrageforschung. *Zeitschrift für Soziologie*, 24, 284-299.
- FUCHS, M. (2001). The impact of technology on interaction in computer-assisted interviews. (Ed. D. W. Maynard, H. Houtkoop-Steenstra, N.C. Schaeffer, H. van der Zouwen). *Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview*. Wiley (à paraître).
- FUCHS, M., COUPER, M., et HANSEN, S. (2000). Technology effects: Do CAPI or PAPI interviews take longer? *Journal of Official Statistics* (sous press).
- GROVES, R.M., et MATTHIOWETZ, N.A. (1984). Computer assisted telephone interviewing: effect on interviewers and respondents. *Public Opinion Quarterly*, 48, 356-369.

cependant trop tôt pour présenter des résultats.

Il y a également lieu d'aborder des questions restées sans réponses à l'aide de recherches plus poussées. Personnellement, j'aimerais proposer une stratégie particulière servant à évaluer ce genre de question à supposer que la conception de questionnaires assistés par ordinateur ait de l'importance pour différents clients: chercheurs, intervieweurs et répondants. Bien entendu, il est important qu'un questionnaire IAO réponde aux besoins du chercheur et que le processus question-réponse soit bien conçu dans chaque cas. Toutefois, à mon avis, il est également important de tenir compte de la dimension sociale de l'interaction intervieweur-répondant et des comportements qui surviennent entre les questions. Si un questionnaire IAO perturbe la dimension sociale du processus de mesure, il risque de nuire à la qualité même des données. Nous ne savons toujours pas quelle stratégie entraîne les meilleurs compromis entre un processus de mesure valable et fiable d'une part, et une interview qui se déroule doucement, brièvement et sans embarras d'autre part. Afin de déterminer dans quelle mesure une conception particulière des écrans IAO pourrait nuire à la qualité des données et comment elle pourrait aider à épargner du temps, de l'argent et des efforts de la part de l'intervieweur, nous devons mener d'autres études d'utilisation.

Afin d'évaluer les questions mentionnées ci-dessus, nous avons vraiment besoin d'autres expériences sur le terrain. Puisque nous voulons analyser la dimension sociale de l'interview et ses effets sur le comportement de l'intervieweur de même que sur la durée de l'interview, les expériences de laboratoire ne répondent pas à tous nos besoins. Bien sûr, les expériences de laboratoire se font dans un contexte plus contrôlé, fournissent des renseignements plus détaillés au sujet des participants et, par conséquent, exigent un moins grand nombre de cas. Et pourtant, sans essais sur le terrain, nous ne confrontons jamais nos prototypes et nos solutions en fonction de la véritable interaction intervieweur-répondant et aussi du processus question-réponse. On devrait donc étudier la

Même si l'effet de la conception quadrillée sur le comportement de l'intervieweur et sur celui du répondant n'est pas considérable, il aide à cerner deux raisons du meilleur déroulement de la version grille/thème pour ce qui est de la durée de l'interview: 1) pour la version grille/thème, l'intervieweur sait bien que le répondant s'adapte mieux à la logique du processus question-réponse, les deux pouvant anticiper la prochaine question plus aisément et le processus question-réponse se déroulant plus doucement; 2) cette version entraîne plus de cas dans lesquels le répondant fournit des renseignements plus rapidement pour les personnes du ménage, et fournit plus souvent des renseignements pour tous les membres du ménage ou du moins pour plusieurs membres en même temps. Les résultats ne sont pas tout à fait uniformes, mais cette version particulière permet à l'intervieweur de s'adapter plus facilement à la situation, d'inscrire les renseignements et d'aider le répondant à fournir la prochaine réponse appropriée sans répétition du texte complet de la question.

Nos résultats se rapportent à la discussion sur la conception des questionnaires d'enquête pour la collecte des données assistée par ordinateur avec la participation d'un intervieweur. Les résultats décrits dans le présent exposé permettent de conclure que le recours à des grilles facilite l'interaction intervieweur-répondant et accélère la collecte des données. Les expériences que nous avons menées, dans le laboratoire de facilité d'utilisation du Survey Research Center de l'Université du Michigan sur des modèles quadrillés et thématiques, indiquent qu'il est possible d'améliorer le rendement des intervieweurs en leur fournissant des grilles (Couper et coll. 1997). Moore et Mayer (1998a; 1998b) ont montré qu'il est possible d'améliorer l'efficacité des interviews en adoptant également un ordre des questions fondé sur un thème. Le présent exposé indique qu'une combinaison des deux fonctions est encore plus avantageuse pour les interviews.

Le recours à des grilles et à un ordre des questions fondé sur un thème suscite un nombre accru de cas dans lesquels l'intervieweur s'éloigne de l'interview standard. D'un point de vue strictement méthodologique, cela peut représenter un sérieux inconvénient, surtout si l'intervieweur s'éloigne de la séquence d'interview standard en ayant recours à des questions globales pour tous les membres du ménage. Ainsi, Martin (1999) a décelé un accroissement appréciable du nombre de personnes dénombrées dans un ménage lorsque l'on pose d'autres questions. De plus, Kindermann et coll. (1997) ont montré, pour des questions fondées sur une liste hors-ménage, que des signaux supplémentaires sur la victimisation entraînent une augmentation appréciable des crimes dénombrés. Si l'on généralise ces résultats pour des questions globales d'une personne à l'autre, on s'attendrait à une diminution de la qualité des données lorsque les intervieweurs ont recours à des comportements non prévus consistant à appliquer des questions globales. Toutefois, les résultats signalés dans le présent exposé indiquent pas que les intervieweurs ont recours à des

questions globales, et l'auteur ne recommande pas d'avoir recours à des questions globales lors de la conception d'un questionnaire d'enquête. Les résultats favorisent plutôt une conception des écrans permettant à l'intervieweur d'utiliser les renseignements fournis par le répondant lorsque celui-ci adopte un mode global et fournit des renseignements pour plusieurs membres du ménage en même temps. Il ne s'agit donc pas d'encourager les chercheurs à utiliser des questions globales, et nous ne souhaitons pas que les intervieweurs modifient les questions prévues de façon à poser des questions globales. Toutefois, lorsqu'un répondant fournit plus de renseignements qu'on ne lui en demande, la conception des écrans d'un questionnaire IA/O ne devrait pas empêcher l'intervieweur d'en bénéficier.

Il a été démontré qu'un modèle fondé sur une grille facilite la tâche des intervieweurs à cet égard, car il permet à ceux-ci de rajuster leur comportement en fonction des règles générales de la conversation. Les résultats du codage des comportements indiquent que les intervieweurs s'éloignent fréquemment des procédures d'interview standard. Oksenberg et coll. (1992, 1993) ont affirmé que ces changements reflètent souvent des rajustements apportés par les intervieweurs devant les modalités de la situation de façon à mieux enchaîner avec ce qui précède immédiatement ou de respecter la situation particulière du répondant. Cela s'impose plus précisément lorsque le répondant ne se limite pas à la question demandée, mais fournit des renseignements supplémentaires. Comme l'ont souligné Oksenberg et coll. (1992, 1995), afin de ne pas donner l'impression de ne pas porter attention à la réponse, les intervieweurs inscrivent souvent la réponse eux-mêmes sans poser la question ou ne posent la question qu'en partie. Ils adoptent donc un comportement axé davantage sur le répondant afin de ne pas paraître insensibles. Une conception des écrans fondée sur une grille et un ordre des questions fondé sur un thème permettent à l'intervieweur de respecter ces règles de la conversation et de réagir en fonction de la situation. Cela pourrait être acceptable ou même préférable, en présence de questions factuelles, pourvu que le comportement de l'intervieweur ne nuise pas à la qualité des données (par exemple questions tendancieuses).

Ce qu'il faut faire afin d'améliorer un questionnaire assisté par ordinateur du point de vue de l'interaction intervieweur-répondant: nos données indiquent que la version grille/thème suscite un déroulement particulier de l'interview auquel l'intervieweur et le répondant peuvent s'adapter facilement. Jeff Moore (1996; Moore et Mayer 1998a, 1998b) a indiqué que les intervieweurs préfèrent la version axée sur un thème. Par contre, nous connaissons mal la satisfaction des répondants à l'égard de cet ordre des questions. Il est donc important d'évaluer leur opinion au sujet de la version différente. De plus, nous ne savons pas si cette version correspond à la façon dont l'information est mémorisée par les répondants. Il se peut que les répondants puissent facilement s'adapter à cette version, mais que d'un

axé sur la grille et le modèle axé sur l'élément est appréciablement plus faible (33,6% contre 26,1%), cependant un effet d'interaction appréciable ($p < 0,05$): le recours à un ordre des questions axé sur le thème n'entraîne pas une différence appréciable. Toutefois, une conception fondée sur une grille jumelée à un ordre des questions axé sur un thème fait augmenter le nombre de cas dans lesquels le répondant fournit des renseignements pour tous les membres du ménage en même temps.

Il est remarquable que les résultats diffèrent même pour les deux conceptions des écrans lorsqu'on utilise un ordre des questions axé sur la personne. L'étude a été menée par téléphone, les répondants n'étant pas du tout au courant de la conception des écrans. La seule explication possible est que l'interviseur modifie son comportement en fonction de la conception des écrans, aiguillant le répondant de façon différente. Les répondants aussi bien que les inter-

vieveurs réagissent donc à la conception des écrans et à l'ordre des questions dans le modèle de type grille/personne d'une façon qui facilite l'interaction intervieweur-répondant, ce qui permet à l'interview de se dérouler plus doucement. (Comme nous l'avons vu, les intervieweurs modifient leur comportement même pour un modèle de type grille/personne (tableau 2), mais l'ordre des questions ne suscite pas chez le répondant un comportement équivalent.) Un inconvénient possible de ces comportements intervieweur-répondant pourrait être une baisse de la qualité des données à cause de modifications du processus de la question-réponse prédéterminé; le répondant considère alors la réponse de façon moins intense et rigoureuse. Nous n'observons que de très rares valeurs manquantes pour une question, de sorte que l'analyse de cet indicateur standard de la qualité des données n'est pas efficace. En réalité, nous ne nous attendons pas à une plus forte proportion de valeurs manquantes pour une question dans l'une ou l'autre version. Toutefois, on pourrait s'interroger sur l'homogénéité des réponses fournies par le répondant. Dans une forte

proportion des cas, le répondant entend le texte complet de la question une seule fois et il est possible que le répondant soit moins rigoureux au moment de répondre à la même question pour des membres subséquents du ménage. De plus, le fait de répondre pour tous les membres du ménage en même temps («nous sommes tous arrivés pendant la même année» risque d'accroître l'homogénéité de la réponse et donc de faire baisser la qualité des données. Afin de pouvoir évaluer cet inconvénient possible, nous avons calculé le nombre de catégories de réponse distinctes choisies par le répondant pour un élément particulier et pour tous les membres du ménage (par exemple pour l'année d'arrivée: le répondant 1985, le partenaire 1987, la fille 1987, le fils 1988 = trois catégories de réponse distinctes). Cela devrait indiquer si les répondants qui adoptent le raccourci («nous sommes tous venus pendant la

Tableau 3

Nombre moyen de catégories distinctes (homogénéité) par ménage selon la version

Variable	Grille + Élément		Grille + Élément +		Total
	thème	thème	thème	personne	
Année d'arrivée (19 catégories)	1,2	1,2	1,2	1,3	1,2
	1,3	1,3	1,3	1,3	
Statut (4 catégories)	1,3	1,3	1,3	1,3	1,3
	1,3	1,3	1,3	1,3	

Aucune différence appréciable

Ces résultats ne fournissent qu'une faible indication que le modèle fondé sur une grille ne nuit pas à la qualité des données. Il faudra utiliser d'autres indicateurs standard de la qualité des données comportant des ensembles de données plus grands si l'on veut déterminer dans quelle mesure la qualité des données est affectée. Toutefois, d'après les données disponibles, nous ne pouvons déceler un effet sur la validité des réponses.

5. DISCUSSION ET CONCLUSION

Les résultats de notre comparaison de quatre versions d'une liste des membres d'un ménage (le même libellé des questions étant utilisé pour toutes les versions) montrent que les intervieweurs tout comme les répondants fonctionnent de façon plus efficace dans le cadre d'un modèle de type grille/thème que pour l'une ou l'autre des trois autres versions. Le fait de combiner un schéma des écrans fondé sur une grille et un ordre des questions fondé sur un thème permet de réduire la durée moyenne de 17% à l'ordre des questions, et un tiers environ à la configuration des écrans. Il importe de mentionner que l'effet de schéma des écrans est moins marqué que celui de l'ordre des questions et encore plus faible, comparativement à l'effet sur la durée, pour le comportement de l'intervieweur et pour le comportement du répondant.

de comportements du répondant par le module assisté par ordinateur.

Table 2
Comportement de l'intervieweur et comportement du répondant selon la version

Grille + Élément thème + thème personne	Proportion moyenne d'éléments touchés par un comportement				
	Grille + Élément thème + thème personne	Grille + Élément thème + thème personne	Grille + Élément thème + thème personne	Grille + Élément thème + thème personne	Grille + Élément thème + thème personne
Total	0,36***	0,21	0,34	0,43	0,48
Le répondant fournit l'information pour toutes les personnes du ménage en même temps	0,48	0,43	0,34	0,29,0%	38,2%

***p < 0,001

Afin de distinguer la proportion de cas touchés par un certain comportement du répondant et la proportion moyenne d'éléments par cas (!) touchés par un certain comportement de l'intervieweur, nous avons utilisé des pourcentages pour ceux-là et des décimales pour ceux-ci.

Afin de pouvoir comparer les quatre versions du modèle d'écran en fonction des comportements d'intervieweur ne correspondant pas à l'interview standard, nous avons calculé la proportion des éléments par cas touchés par ce genre de comportement. Les quatre versions manifestent des écarts fort différents relativement à la séquence prévue de ce genre (la proportion moyenne des éléments touchés des interviews: l'application de la version grille/thème à une interview entraîne plus de deux fois plus de comportements de ce genre (la proportion moyenne des éléments touchés de ce genre (0,48) que la version élément/personne (0,21) qui représente la norme pour la plupart des études menées jusqu'à présent. (Une analyse de la variance indique que les deux facteurs influencent l'effet global indépendamment (conception des écrans: p < 0,001; ordre des questions: p < 0,001; aucun effet appréciable d'interaction). On peut attribuer 25% environ de l'effet global à la conception des écrans et 75% aussi le temps consacré à l'interview: les éléments influencent aussi le temps consacré à l'interview qui diffère de touches par un comportement d'intervieweur qui diffère de la séquence prévue de l'interview prennent appréciablement moins de temps (4,0 secondes) que les éléments qui respectent la séquence ordinaire (6,8 secondes; p < 0,001).

2. De plus, une analyse du comportement des répondants indique que le modèle thématique suscite une plus forte proportion de cas (42,3% comparativement à 19,7% pour la stratégie personnelle; p < 0,001) dans lesquels le répondant fournit au moins une fois des renseignements pour toutes les personnes ou pour un groupe de personnes du ménage en même temps pendant la même année»; nous avons tous le même statut juridique». Par contre, la différence entre le modèle

dévier de l'interview prévue de façon plus marquée que les autres versions. Pour ce qui est de la durée, cette version permet à l'intervieweur d'utiliser de façon efficace l'information fournie pour tous les membres du ménage en même temps. L'analyse du codage vidéo appuie notre interprétation de l'occurrence propre à une version de comportement des intervieweurs (1) et de comportements des répondants (2) qui permettent de gagner du temps:

1. L'analyse des bandes vidéo permet d'observer des comportements d'intervieweurs ne correspondant pas aux procédures standard: quelque 78% des éléments sont lus tels quels, et 9,3% des éléments ne sont pas présentés au répondant par l'intervieweur. De plus, dans 5% des cas, l'intervieweur ne lit pas la question, mais fournit plutôt un autre stimulus comportant le lien entre la personne suivante et le répondant (par exemple «... et votre épouse?»). (Il est intéressant de noter que les intervieweurs choisissent les mêmes expressions verbales, spontanément, que Moore et Moyer 1998a, 1998b ont utilisées dans leurs expériences sur l'ordre des questions.) Dans 5,5% des cas, l'intervieweur ne lit pas la question, mais vérifie plutôt la réponse («... et votre épouse a 32 ans?»). On observe également quelques questions incomplètes et des réponses inscrites incorrectement. Au total, 22% environ des éléments sont touchés par au moins un comportement d'intervieweur qui ne correspond pas à une séquence d'interview standard – ce qui représente une valeur étonnamment élevée puisque tous les intervieweurs savaient que les interviews étaient enregistrées sur bande magnétique! Toutefois, comparativement à d'autres études sur le comportement des intervieweurs, les valeurs sont appréciablement moins élevées. Oksenberg, Cannell et Bixt (1996), par exemple, ont appliqué un codage du comportement à la National Medical Expenditure Survey, observant de 37% à 41% de comportements d'intervieweur de ce genre. Il sera question ci-dessous de la mesure dans laquelle ce genre de comportement aide à obtenir des mesures valables.

De ces résultats particuliers, nous tirons la conclusion suivante: la plupart de ces comportements représentent une espèce de raccourci, par exemple l'intervieweur ne lit pas le texte exact de la question, tente de faciliter la conversation ou de la rendre plus conforme aux règles de la conversation. À notre avis, cela indique que l'intervieweur ne veut pas poser une question à laquelle une réponse a déjà été fournie. L'intervieweur ne veut pas rester indifférent devant l'information fournie par le répondant, cherchant plutôt à respecter les règles de la conversation. Ces comportements prennent donc moins de temps que des comportements d'intervieweur standard. Selon nous, la priorité ne consiste donc pas à épargner du temps, mais bien à personnaliser le processus question-réponse en fonction

donne une valeur de 5,5 secondes par élément. (Une analyse de la variance indique que les deux facteurs – c'est-à-dire la conception des écrans et l'ordre des questions – favorisent indépendamment la diminution du temps (conception des écrans: $p > 0,01$, un tiers de l'effet total; ordre des questions: $p > 0,001$, deux tiers de l'effet total, aucune interaction appréciable).)

Pourquoi la version grille/thème est-elle la plus rapide? Une analyse détaillée indique que cette version est particulièrement rapide pour la collecte des renseignements sur la deuxième personne et les personnes subséquentes du ménage – effet appréciable que l'on appelle l'effet de boucle (Fuchs 2001). Cette expression décrit le phénomène suivant: l'intervieweur consacre beaucoup plus de temps à recueillir les renseignements pour la première personne du ménage comparativement aux personnes subséquentes. L'effet de boucle moyen donne au total 3,4 secondes par élément, ce qui représente une réduction de 38% environ relativement à la première personne (tableau 1).

L'effet de boucle n'est pas particulièrement propre à cette expérience. Nous avons observé des effets de boucle dans nos expériences antérieures liées à la liste des membres des ménages NHIS (Couper et coll. 1997). Toutefois, il est intéressant de noter que l'effet de boucle est apprécié beaucoup plus grand pour les versions axées sur le thème qu'il ne l'est pour les versions qui suivent un ordre des questions axé sur la personne (tableau 1). Ainsi, les versions fondées sur le thème permettent d'accélérer le processus pour la deuxième personne et les personnes subséquentes d'un ménage, et ce comportement a un effet de boucle plus grand. (Il est possible que la conception de notre expérience a eu comme effet que des intervieweurs ne savaient pas quelle version ils utilisaient, ce qui a pu faire baisser leur rendement pour le tout premier élément. Or cet effet devrait être le même pour toutes les versions, de sorte que les résultats ne devraient pas être touchés.)

Tableau 1

Durée et effet de boucle (en secondes)

Elément	Âge	Arrivée	Situation	Tous les éléments
Première personne du ménage	9,4	9,9	7,7	9,0
Toutes les autres personnes du ménage	6,6	5,7	4,5	5,6
Toutes les personnes	8,0***	7,8***	6,1***	7,2***
Différence entre la première et toutes les autres personnes du ménage				
Effet de boucle				
Grille + thème	-6,8	-6,4	-4,6	-5,9
Elément + thème	-5,2	-8,3	-4,3	-6,0
Grille + personne	-0,5	-2,7	-3,0	-2,1
Elément + personne	0,3	-0,3	-1,3	-0,4
Effet de boucle moyen	-2,8***	-4,2***	-3,2**	-3,4***

L'analyse des bandes vidéo nous a permis de fournir des raisons partielles au moins pour ces différences: dans des conditions axées sur le thème, les intervieweurs et les répondants s'adaptent différemment à l'interview comparativement aux versions fondées sur la personne. Lorsque l'on pose les questions pour toutes les personnes du ménage, le répondant reconnaît très rapidement le caractère logique du processus. Dans une forte proportion des cas (30% environ), la réaction est du type: «nous sommes tous arrivés pendant la même année» (ce qui veut dire: «cessez de me répéter cette question»).

Si le questionnaire suit la conception fondée sur la personne, l'intervieweur doit mémoriser cette information, et au moment d'aborder la personne suivante, il doit se rappeler: «ne répète pas cette question, car le répondant a déjà fourni la bonne réponse!». Parfois l'intervieweur s'en souvient, mais la plupart du temps il pose la question de nouveau. Cela est particulièrement vrai pour une configuration d'écran axée sur l'élément et ne fournissant aucune indication des réponses à la même question pour les autres membres du ménage. Dans un modèle fondé sur le thème, par contre, l'intervieweur peut facilement s'adapter à cette situation. Il suffit alors d'entrer le même code pour toutes les personnes du ménage sans répéter la question. L'intervieweur aussi bien que le répondant s'habituent aux questions, de sorte que le processus question-réponse comporte moins d'interventions verbales de la part de l'intervieweur comme de celle du répondant. L'intervieweur aussi bien que le répondant peuvent anticiper la prochaine question, et l'interview se déroule plus facilement. Cela est particulièrement vrai lorsque le module LAO comporte une grille et fournit d'autres informations contextuelles, par exemple, les réponses fournies pour d'autres membres du ménage à la même question. (Les résultats figurant dans la partie inférieure de la figure 4 permettent de conclure que la version grille/personne ne bénéficie pas dans la même mesure des avantages de la stratégie fondée sur le thème. Toutefois, à cause de la conception de la grille, l'effet de boucle est appréciablement plus grand que pour la version élément/personne). Par conséquent, le temps consacré par élément est appréciablement plus court, et l'intervieweur peut adopter un comportement orienté vers le répondant à celui dont il est question dans Schöber et Conrad (1997).

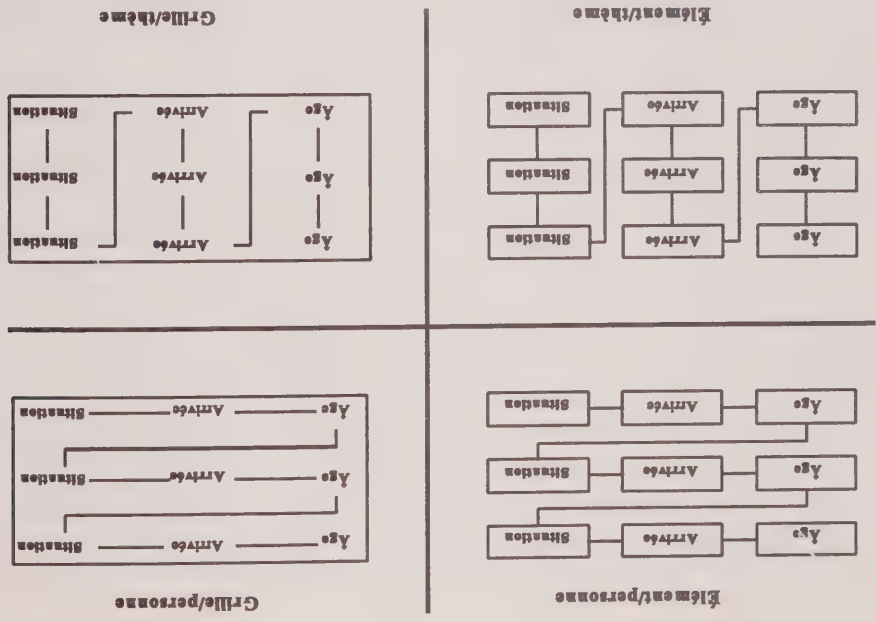
L'inclusion de remarques par l'intervieweur sert parfois de signal que la réponse à la question précédente a été inscrite de façon à aiguiller le répondant, celui-ci devenant la question suivante et fournissant la réponse appropriée même sans stimulation supplémentaire. Dans un cas extrême, cela peut mener le répondant à fournir l'information au sujet de toutes les personnes du ménage en même temps: «nous sommes tous venus pendant la même année». Les diverses versions mises à l'essai dans cette expérience favorisent ce genre de comportement à différents degrés. Les résultats nous permettent de conclure que la version grille/thème pousse les intervieweurs et les répondants à

La durée a été appréciablement différente pour chacune des quatre versions: les intervieweurs ont consacré 6,6 secondes par élément dans la version élément/personne

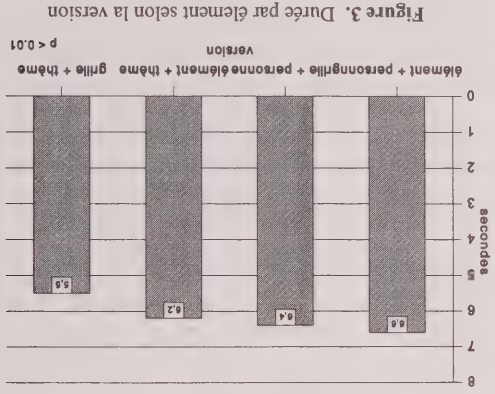
4. RÉSULTATS

Quatre versions d'une liste des membres d'un petit ménage, avec trois éléments par personne, ont été incluses dans le module: une version élément/personne, une version grille/thème. Toutes les versions ont fait appel au même grille/thème. L'ordre des questions et aux mêmes instructions pour l'intervieweur, mais nous avons modifié la conception des écrans et l'ordre des questions en fonction de la stratégie théorique considérée comme la version standard – elle représente la conception des questionnaires appliquée habituellement aux volets socio-démographiques des enquêtes IAO. Une des quatre versions a été attribuée de façon aléatoire à chaque interview – et donc aux intervieweurs et aux répondants. Nous avons mesuré le temps total requis pour la liste des membres du ménage, ainsi que le temps consacré à chaque élément de cette section des 501 interviews. De plus, 234 interviews ont été choisies au hasard, et nous avons enregistré sur bande magnétoscopique l'intervieweur en train de parcourir la section de la liste des membres du ménage. Les segments vidéo ont été codés en fonction du comportement de l'intervieweur et du répondant, et les données résultantes ont été combinées aux mesures du temps.

Figure 2. Quatre versions vérifiées dans l'expérience (chaque cases représente un écran)



Il importe de mentionner que les deux facteurs semblent jouer un rôle dans la diminution du temps mis à accomplir la tâche. Si nous distinguons les deux facteurs, nous obtenons les résultats suivants: les deux versions axées sur le thème sont appréciablement plus courtes que les deux versions fondées sur la grille supposent un temps appréciablement moindre que les deux versions axées sur la personne. L'effet combiné s'applique à la version grille/thème, et



(considérée comme la version standard). Par contre, chaque élément a supposé 5,5 secondes dans la version grille/thème. C'est là une réduction de 17% environ pour la version grille/thème. Les deux autres versions se situent entre les deux premières.

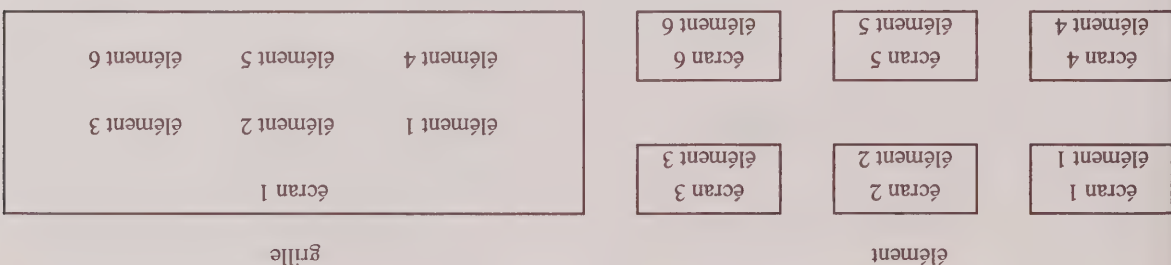


Figure 1. Conception fondée sur l'élément et conception fondée sur la grille

modèle thématique comporte une non-réponse moins élevée, moins d'interruptions et de refus, et il prend appéciablement moins de temps. Les intervieweurs préfèrent nettement ce modèle.

fait appel aux avantages de la stratégie thématique et de la conception des écrans fondée sur une grille: nous avons combiné les deux modèles d'écran (conception fondée sur l'élément et conception fondée sur la grille) et les deux ordres des questions (ordre axé sur la personne et ordre axé sur le thème) de façon à vérifier les quatre versions résultantes dans une expérience sur le terrain. Pour ce faire, nous avons adopté l'hypothèse qui suit: la facilité d'utilisation d'un module IAO ne relève pas seulement de la programmation, mais elle est liée à la conception du questionnaire et au contexte social de l'interview. Les deux aspects d'un module assisté par ordinateur, c'est-à-dire la conception des écrans et l'ordre des questions, peuvent favoriser ou détruire la bonne marche d'une interview. Les résultats de la recherche antérieure avaient donné lieu à l'hypothèse suivante: la combinaison d'une conception des écrans axée sur une grille et d'un ordre des questions axé sur un thème suscite l'interaction la plus efficace entre l'intervieweur et le répondant.

3. METHODE

L'expérience s'est déroulée en Allemagne en mars 1998. On a interrogé des immigrants d'origine allemande provenant de la Pologne, de la Roumanie et de l'ancienne Union soviétique. À compter du 28 février 1998, et jusqu'au 20 mars 1998, les intervieweurs ont mené $n = 501$ interviews. On a fait parvenir aux répondants une lettre d'introduction à l'avance, et on a communiqué avec eux par téléphone jusqu'à 15 fois. Le taux de réponse a été de 84%, et la non-réponse à une question a été assez faible. Les interviews ont été menées dans le cadre du programme ITAO C13. Quelque 95 questions ont été posées sur divers thèmes. En moyenne, les interviews ont duré 23 minutes.

référence menée sur deux ordres des questions servant à recueillir des informations sur toutes les personnes admissibles d'un ménage (Moore et Moyer 1998a, 1998b). Selon le premier modèle, on pose toutes les questions pour la première personne admissible du ménage, puis on passe à la personne suivante. Un tel ordre des questions est ce que l'on appelle la stratégie axée sur la personne. Selon le deuxième modèle, appelée la stratégie thématique, la première question est posée pour toutes les personnes admissibles, puis on pose la deuxième question pour toutes les personnes, et ainsi de suite. Les résultats de Moore et Moyer appuient fortement une conception thématique: le

(Oksenberg, Beebe, Blixt et Cannell 1992). En présence d'un questionnaire imprimé, il est facile pour un intervieweur d'utiliser immédiatement l'information supplémentaire fournie par le répondant. Pour une réponse du type « nous sommes tous noirs », par exemple, l'intervieweur peut facilement cocher les cases appropriées pour tous les membres du ménage en même temps. Une personne qui s'intéresse à la conception des questionnaires peut alors se poser la question suivante : vu le manque de souplesse d'un environnement assisté par ordinateur, quel est le meilleur ordre des questions si l'on veut recueillir des renseignements pour tous les membres du ménage ?

sur la personne», voir Couper et coll. 1997 et Fuchs 2001 ; au sujet des «questions groupées», voir Mayer (1996). Ainsi, le module IAO peut exiger que l'on demande l'âge du répondant, son niveau de scolarité et d'autres questions avant de passer à l'âge de son épouse. (Cela s'explique en partie par le fonctionnement des programmes informatiques et des bases de données : les ménages représentent les principales catégories, les personnes ou les autres entités étant traitées comme des sous-catégories.) Lorsqu'on remplit le questionnaire pour une liste de membres du ménage, il arrive (et même assez souvent, comme on le verra ci-dessous) que le répondant fournisse non seulement la réponse à la question posée (par exemple «j'ai 34 ans») mais également à une question connexe : «j'ai 34 ans et mon épouse en a 32» ; le répondant pourra aussi déclarer : «nous sommes tous noirs» lorsqu'il est question de race

durée des interviews et de la facilité d'utilisation (on trouvera une synthèse dans Couper et coll. 1998). Le présent document fait partie de cette récente discussion des «effets de la technologie» (Fuchs, Couper et Hansen 2000).

2. CONTEXTE THÉORIQUE

Dans la présente analyse, le point de convergence théorique touche surtout à deux aspects de la facilité d'utilisation: 1) la segmentation du cheminement de l'interview et 2) le manque de souplesse pour l'intervieweur.

1. Segmentation: Dans un environnement IPAO, l'intervieweur se voit imposer un fardeau supplémentaire: le processus de saisie a lieu durant l'interview. Normalement, un intervieweur lit une question, reçoit une réponse, entre les données, appuie sur [enter] et voit ensuite apparaître le prochain écran et la question suivante. Comparativement à l'IPC, les intervieweurs ne peuvent pas regarder en avant et anticiper la question suivante tout en inscrivant les réponses à la question précédente, et ils ne peuvent pas commencer la question précédente, et ils ne peuvent pas commencer à lire la question suivante avant d'appuyer sur [enter] – ils ne peuvent pas s'occuper des deux tâches en même temps. Dans ce genre de procédure, l'interaction entre l'intervieweur et le répondant est segmentée par des touches [enter]. Il n'existe encore aucune indication quantitative que ce genre de segmentation nuit aux données ou au déroulement de l'interview. Par contre, on fait remarquer que l'intervieweur n'a pas de «vue d'ensemble» et que la pertinence des questions et les relations entre elles ne sont peut-être pas claires (House 1985; Groves et Mathiowetz 1984).

Nos résultats, obtenus de plusieurs séries d'essais de facilité d'utilisation menées en laboratoire sur la configuration des écrans d'une liste de membres de ménage (Couper et coll. 1997; Hansen, Couper and Fuchs 1998), ont mis en valeur une configuration d'écran particulière qui permet à l'intervieweur de mieux comprendre le questionnaire, tout en conservant l'interaction avec le répondant et tout en inscrivant les données. Deux versions d'une série de questions ont été mises à l'essai en laboratoire en fonction du temps mis à remplir le questionnaire et de la facilité d'utilisation. Nous avons comparé une conception dite thématique et une conception dite quadrillée. House et Nichols (1988) ont décrit trois stratégies de conception des écrans pour des modules assistés par ordinateur, axée sur l'élément, axée sur l'écran et axée sur la forme. Un module axé sur l'élément fait afficher une question et un champ de saisie à la fois, et les opérations logiques ont lieu durant la transition d'un écran à l'autre. Facile à programmer, une telle conception dirige l'attention de l'intervieweur sur la question même. Un module axé sur l'écran regroupe plusieurs éléments supposant des réponses séquentielles. Toutes les opérations logiques sont exécutées après chaque

2. Manque de souplesse: La deuxième caractéristique qui risque de poser des problèmes dans une interview assistée par ordinateur est le manque de souplesse. Un IAO peuvent faire amplement usage de l'enchâinement peut guère sauter de questions. Même si les modules IAO peuvent faire amplement usage de l'enchâinement des questions et de filtres, l'ordre des questions est déterminé d'avance. Normalement, il faut appuyer sur la touche [enter] pour chaque question avant que le système ne passe à l'écran suivant. On trouve cette rigidité de l'ordre des questions avantageuse car elle évite toute difficulté de la part de l'intervieweur de suivre le déroulement, les questions destinées à des répondants particuliers, les filtres, l'enchâinement des questions, etc. L'intervieweur n'a pas à s'en soucier et il en résulte un ordre très rigide des questions et très peu de souplesses, pour l'intervieweur, quant à l'ordre des questions. Voici un exemple qui illustre bien cet effet: la plupart des modules IAO appliquent un ordre des questions à la liste des membres du ménage, toutes les questions adressées à une personne étant posées avant que l'intervieweur ne pose les mêmes questions à la personne suivante (au sujet de la «conception axée

Nous avons observé que la conception fondée sur la grille permet de réduire la segmentation. Les intervieweurs peuvent commencer à lire la question suivante tout en inscrivant les données de la question précédente. Même un recul en arrière semble de plus facile dans la conception fondée sur une grille. Par contre, nous n'avons constaté qu'un faible appui pour la conception fondée sur une grille relativement au temps mis à exécuter la tâche (on trouvera des détails dans Couper et coll. 1997). On peut donc poser la question: Que faire pour réduire la segmentation et améliorer davantage l'efficacité d'une liste de membres de ménage pour ce qui est de la durée?

La version axée sur l'élément vérifiée dans notre expérience correspond aux caractéristiques décrites par House et Nichols (1988) pour une stratégie fondée sur l'écran. Pour sa part, la conception fondée sur une grille est en réalité un module axé sur la forme. Elle permet aux intervieweurs d'enregistrer l'information dans l'ordre choisi par le répondant; elle donne à l'intervieweur une meilleure vue d'ensemble du module, et les mises à jour ainsi que les reculs en arrière sont plus faciles (on trouvera des détails dans Couper et coll. 1997). De plus, elle permet à l'intervieweur d'avoir plus d'une question à un même écran (avantages: vivesse d'exécution et informations contextuelles). Le graphique ci-dessous illustre la conception des écrans fondée sur l'élément et celle fondée sur la grille pour les IAO.

La conception des écrans et l'ordre des questions dans un module IAO

Résultats d'une expérience de facilité d'utilisation menée sur le terrain

MARK FUCHS¹

RÉSUMÉ

La conception des écrans et la conception des questionnaires ont une influence sur le comportement des intervieweurs dans un environnement IAO. Des recherches antérieures ont montré que les intervieweurs peuvent travailler de façon plus convenable et efficace si des fonctions et caractéristiques appropriées sont incorporées dans le module IAO. Des expériences fondées sur la liste des membres des ménages de deux importantes enquêtes gouvernementales ont montré que le recours à des grilles et à des tableaux facilite le travail de l'intervieweur. Ces expériences ont été menées dans des conditions de laboratoire, mais nous avons des résultats d'une première expérience sur le terrain. En mars 1998, une enquête de type ITAO a été menée auprès d'immigrants en Allemagne (taux de réponse 84%, n = 501). Dans cette étude de production, quatre versions d'une liste des membres du ménage ont été comparées afin de vérifier deux conceptions des écrans et deux ordres des questions dans une conception factorielle 2x2. Les quatre versions ont été attribuées de façon aléatoire à des intervieweurs et à des répondants. Des mesures du temps ont été intégrées au programme ITAO, et un schéma de codage. À l'aide des données, nous avons évalué la facilité d'utilisation des différentes caractéristiques de la conception de l'IAO. Les résultats indiquent que la conception des écrans de même que l'ordre des questions ont une influence appréciable sur la durée de l'interview et le comportement des intervieweurs. En particulier, la version quadrillée et thématique donne les résultats les plus rapides pour ce qui est du temps mis à remplir le questionnaire. Les résultats tirés des données de codage indiquent que les différences entre les versions sont attribuables à des comportements particuliers des intervieweurs et des répondants. Les données indiquent que la version thématique quadrillée facilite chez l'intervieweur un comportement axé sur le répondant et donne lieu au meilleur rendement des intervieweurs pour ce qui est de la durée.

MOTS CLÉS: Interview assistée par ordinateur; essai de facilité d'utilisation; expérience sur le terrain; conception des écrans; ordre des questions.

1. INTRODUCTION

Les interviews assistées par ordinateur sont en train de devenir une technique d'enquête standard (Couper, Baker, Bethlehem, Clark, Martin, Nicholls et O'Reilly 1998). Pour les enquêtes téléphoniques aussi bien que pour les interviews personnelles, de plus en plus d'études sont menées à l'aide de techniques d'interview assistée par ordinateur (IAO). Plusieurs enquêtes gouvernementales importantes aux États-Unis sont en train de passer à l'IAO, ou cette transition est déjà terminée. Même en Europe, on observe une évolution vers l'interview assistée par ordinateur (Schneid 1991; Fuchs 1994, 1995; Laurie et Moon 1997; Projektgruppe SOEP 1998) – même si, pour le moment, les aspects méthodologiques de ce phénomène ne représentent pas le point central de la recherche européenne. Les chercheurs et les responsables des enquêtes se fient à l'interview assistée par ordinateur pour plusieurs raisons (toutefois, il semble parfois que de solides arguments sont moins importants que la simple popularité de l'IAO): ils espèrent recueillir des données de qualité supérieure grâce à des vérifications incorporées de la cohérence et des intervalles au cours de l'interview;

- l'IAO permet d'enchaîner les questions automatiquement et de concevoir des questionnaires plus complexes sans imposer un fardeau indu à l'intervieweur;
 - ils espèrent consacrer moins de temps et d'argent à l'interview et au post-traitement et réduire les budgets d'enquête une fois les coûts initiaux de matériel et logiciel absorbés;
 - ils espèrent bénéficier de la capacité de l'IAO d'intégrer des données externes à l'interview, fonction particulièrement intéressante pour les études par panel.
- La transition générale à l'IAO est vue d'un bon œil. Les chercheurs et les directeurs sur le terrain en tirent des avantages (Nicholls et de Leeuw 1996) et les intervieweurs (Couper et Burt 1994) de même que les répondants (Baker 1992) manifestent beaucoup de sympathie ou du moins de l'acceptation. Par contre, les interviews assistées par ordinateur ont introduit certains problèmes supplémentaires dans le déroulement des interviews: au début, la recherche méthodologique a porté surtout sur des problèmes de matériel et de logiciel (on trouvera des synthèses dans Couper, Groves et Kosary 1989 et dans Weeks 1992). Des études plus récentes ont traité de l'acceptation par les intervieweurs et les répondants, de la

- SCHUMAN, H., et PRESSER, S. (1981). *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording and Context*. New York: Academic Press.
- SCHWARZ, N., et SUDMAN, S. (1995). *Answering Questions*. San Francisco: Jossey-Bass.
- SNIJDER, T. (1996). Analysis of longitudinal data using the hierarchical linear model. *Quality & Quantity*, 30, 405-426.
- SNIJDER, T., et BOSKER, R. (1999). *Multi-level Analysis. An introduction to basic and advanced multi-level modeling*. London: Sage Publications.
- YANG, M., et GOLDSTEIN, H. (1996). Multilevel models for longitudinal data. Dans *Analysis of Change. Advanced Techniques in Panel Data Analysis*. (Eds. U. Engel, et J. Reinecke). Berlin - New York: Walter de Gruyter, 191-220.
- SUDMAN, S., et BRADBURN, N. (1974). *Response Effects in Surveys*. Chicago: Aldine.



Graphique 3. Résidus standardisés liés aux répondants selon les résidus standardisés liés aux intervieweurs.

De nouveau, ce graphique attire l'attention sur le plus grand nombre d'écarts vers le haut et sur les valeurs extrêmes liées aux intervieweurs. Ces déviations mises à part, on n'observe aucune régularité particulière. À cause des valeurs extrêmes liées aux intervieweurs, le nombre d'observations est plus faible du côté droit que du côté gauche du graphique. Cependant, les résidus liés aux répondants n'ont pas vraiment tendance à être plus faibles si les résidus liés aux intervieweurs sont plus élevés. Le graphique ne permet pas non plus de supposer que l'inverse

est vrai. La vérification présentée au graphique 3 est imparfaite, car elle consiste à attribuer aux répondants les résidus liés aux intervieweurs. Une meilleure solution consisterait à ajuster un modèle plus complexe tenant compte d'une interaction entre les deux effets aléatoires. Goldstein (1995, p. 119) propose ce genre de modèle. L'exécution d'un test en vue de déterminer l'amélioration du modèle grâce au terme d'interaction donne une idée de l'existence d'une correction éventuelle entre les résidus. Une autre solution consisterait à ajouter un niveau supplémentaire (la région) aux niveaux des intervieweurs et des répondants. Ce modèle inclurait un terme pour la variation régionale qui entraînerait une corrélation entre les résidus liés à l'intervieweur et au répondant. Snijders et Bosker (1999, p. 159-160) décrivent ce modèle. Toutefois, les deux modèles nécessitent un paramétrage différent avec création d'ensembles distincts de variables fictives. Leur explication ferait, en soi, l'objet d'un rapport et dépasse donc le cadre du présent article.

BIBLIOGRAPHIE

- BARTLAR, B., BAILEY, L., et STEVENS, J. (1977). Measures of interviewer bias and variance. *Journal of Marketing Research*, 14, 337-343.
- RAUDENBUSH, S. W. (1993). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational Statistics*, 18, 321-349.
- SÄRNDA, C., SWENSSON, B., et WRETTMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- BEERTEN, R., BILLIET, J., CARTON, A., et SWYNGEDOUW, M. (1997). *1995 General Election Study Flanders-Belgium. Codebook and Questionnaire*. Leuven: ISPO/Département Sociologie K.U.Leuven.
- BRYK, A. S., et RAUDENBUSH, S. (1992). *Hierarchical Linear Models Applications and Data Analysis Methods*. Newbury Park - London: Sage.
- CARTON, A., SWYNGEDOUW, M., BILLIET, J., et BEERTEN, R. (1993). *Source Book of the Voters' Study in Connection with the 1991 General Election*. Leuven: Sociologisch Onderzoeksinstituut/ISPO.
- DIPRETE, T. A., et FORRISTAT, J. D. (1994). Multilevel models: Methods and substance. *Annual Review of Sociology*, 20, 331-357.
- GOLDSTEIN, H. (1995). *Multilevel Statistical Models*. London: Edward Arnold.
- GROVES, R. M. (1989). *Survey Error and Survey Costs*. New York: Wiley.
- HANSON, R. H., et MARKS, E. S. (1958). Influence of the interviewer on the accuracy of survey results. *Journal of the American Statistical Association*, 53, 635-655.
- HOX, J. J. (1994). Hierarchical regression models for interviewer and respondent effects. *Sociological Methods and Research*, 22, 300-318.
- HOX, J. J., DE LEEUW, E. D., et KREFT, I. G. (1991). The effect of interviewer and respondent characteristics on the quality of survey data: a multilevel model. Dans *Measurement Errors in Surveys*. (Éds. P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, et S. Sudman). New York: Wiley, 439-461.
- ISPO/PIOP (1995). *1991 General Election Study Belgium. Codebook and Questionnaire*. Leuven: ISPO.
- KREFT, I. G., et DE LEEUW, J. (1998). *Introducing Multilevel Modeling*. London: Sage Publications.
- KROSNICK, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5, 213-236.
- LOOSEVELDT, G., et CARTON, A. (1997). Evaluation of nonresponse in the Belgian Election Panel Study '91 - '95. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1017-1022.
- RASBACH, J., et GOLDSTEIN, H. (1994). Efficient analysis of mixed hierarchical and cross-classified random structures using a multilevel model. *Journal of Educational Statistics*, 19, 337-350.
- RASBACH, J., et WOODHOUSE, G. (1996). *MLn Command Reference*. Version 1.0a. London: Multilevel Models Project. Institute of Education, University of London.
- RAUDENBUSH, S. W. (1993). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational Statistics*, 18, 321-349.

ANNEXE 2

À la quatrième section, nous décrivons les hypothèses sur lesquelles s'appuient les divers modèles utilisés. Pour le dernier, les hypothèses les plus importantes ont trait aux effets aléatoires associés au répondant et à l'intervieweur. On peut évaluer l'hypothèse selon laquelle les valeurs de $\sigma^2_{\text{constant}}$ obtenues pour le répondant et pour l'intervieweur suivent la loi normale en examinant le graphique 1 présentant la courbe obtenue pour les résidus standardisés liés aux répondants et le graphique 2, pour les résidus standardisés liés aux intervieweurs.

Les conclusions que l'on peut tirer de ces graphiques sont les suivantes: les résidus ne présentent aucune anomalie manifeste, mais il pourrait être utile de poursuivre l'étude pour déterminer la raison des déviations plus nombreuses vers le haut et des valeurs extrêmes pour les résidus relatifs à l'intervieweur. À l'heure actuelle, on ne dispose pas de méthodes efficaces pour faire ces vérifications dans le cas des modèles multilivreaux (Goldstein 1995 p. 29). Mais on peut évidemment analyser l'ensemble de données sans les valeurs extrêmes. C'est ce que nous avons fait au tableau 4.

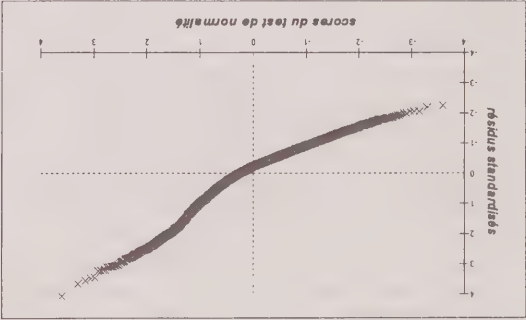
Tableau 4

Analyse de l'ensemble de données sans les valeurs extrêmes liés aux intervieweurs (é.-l. entre parenthèses)

Fixe			
Niveau des mesures	constante	3,853	(0,162)
Niveau des répondants	sexe	1,820	(0,153)
	niveau de scolarité	-1,929	(0,149)
	pressel	1,160	(0,090)
	pressel2	1,217	(0,102)
Aléatoire			
Niveau 2	Intervieweurs	$\sigma^2_{\text{constant}}$	2,495 (0,333)
	Répondants	$\sigma^2_{\text{constant}}$	7,109 (0,530)
	Niveau des mesures	$\sigma^2_{\text{constant}}$	13,420 (0,481)
-2 LL			26850,2

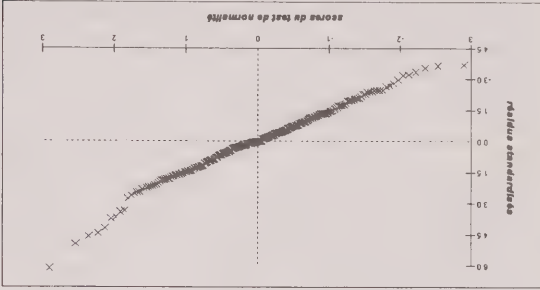
Pour procéder à l'analyse présentée au tableau 4, nous avons exclu deux intervieweurs, à savoir celui qui correspond à la valeur résiduelle la plus faible et celui qui pond à la valeur résiduelle la plus élevée. Les coefficients qui figurent dans ce tableau sont fort semblables à ceux présentés pour le modèle c au tableau 3. La variance liée à l'intervieweur a diminué légèrement, grâce à l'exclusion des deux valeurs extrêmes, mais les résultats ne donnent aucune preuve que les valeurs extrêmes ont un effet important sur les résultats.

L'autre hypothèse concernant les effets aléatoires associés à l'intervieweur et au répondant est celle qui a trait à leur indépendance réciproque. Les résidus liés à l'intervieweur et au répondant ne devraient pas être corrélés. Naturellement, cette hypothèse est plus difficile à vérifier, parce que les résidus sont chacun lié à leur unité respective et que ces unités ne correspondent pas. On obtient 3 028 résidus liés aux répondants et 275 liés aux intervieweurs. Il est possible de vérifier indirectement cette hypothèse en attribuant aux répondants les résidus liés aux intervieweurs. C'est ce qu'illustre le graphique 3.



Graphique 1. Résidus standardisés liés aux répondants selon le score d'équivalence à la loi normale

Dans ce graphique, les écarts par rapport à la diagonale sont assez limités et ne permettent d'inférer aucune violation de la loi normale. Par ailleurs, il convient de souligner que le nombre d'observations est plus élevé dans le cadran supérieur droit que dans le cadran inférieur gauche. Sur le graphique 2, on n'observe pas non plus d'écart net par rapport à la diagonale. Toutefois, dans ce graphique, certaines valeurs extrêmes attirent l'attention, particulièrement celles situées dans le cadran supérieur droit, qui semblent tomber en-dehors de la fourchette des autres résidus liés aux intervieweurs. De surcroît, le graphique contient lui aussi un plus grand nombre d'observations dans le cadran supérieur droit que dans le cadran inférieur gauche.



Graphique 2. Résidus standardisés liés aux intervieweurs selon le score d'équivalence à la loi normale.

comme le montre notre analyse. La souplesse de ce modèle plus que compense l'impossibilité d'y inclure les interactions entre répondant et intervieweur. On pourrait inclure ces interactions si l'on analysait séparément chaque cycle de l'enquête par panel. Toutefois, ces analyses ne permettraient pas de modéliser une évolution éventuelle de la variable dépendante, un autre avantage important de l'analyse conjointe de tous les cycles de l'enquête par panel.

REMERCIEMENTS

Nous remercions le centre de recherches electorales ISPO-PIOP de nous avoir fourni les données. Jacques Billier, Marc Swyngedouw, Ann Carton et Roeland Beerten ont recueilli les données originales sur la région flamande. L'ISPO-PIOP est financé par les Services fédéraux des affaires scientifiques, techniques et culturelles. Ni les personnes qui ont recueilli les données au départ ni le centre de recherche ne sont responsables des analyses ou des interprétations présentées ici. Nous remercions également Jon Rasbash (Institute of Education, University of London) pour certains commentaires fort utiles concernant l'exploitation du logiciel *MLn*. Enfin, nous remercions les examinateurs pour leur suggestions et commentaires constructifs concernant une version antérieure de l'article.

ANNEXE 1

La question était la suivante: «On dit que les partis politiques sont «catholiques» ou «non catholiques». Veuillez placer la carte de chaque parti sur la case de la carte numéro 20 qui correspond le mieux à la mesure dans laquelle le parti en question est «catholique» ou «non catholique». Si, selon vous, deux ou plusieurs partis sont aussi «catholiques» ou aussi «non catholiques», placez les cartes sur la même case. Si vous ne savez pas dans quelle mesure un parti est «catholique» ou «non catholique», mettez simplement la carte de côté.» Avec la carte:

Catholique 0 1 2 3 4 5 6 7 8 9 10 non catholique

La question sur le suivi de la politique dans la presse n'était pas la même pour les deux cycles de l'enquête. Pour le premier cycle, la question concernant le suivi dans la presse était: «À quelle fréquence lisez-vous les nouvelles politiques dans les journaux?» Les catégories de réponses étaient les suivantes:

1=(presque) toujours; 2=souvent; 3=de temps en temps; 4=rarement; 5=jamais.

Lors du deuxième cycle, la question est devenue: «À quelle fréquence suivez-vous les nouvelles politiques à la radio, à la télévision ou dans les journaux?» Les catégories de réponses sont restées les mêmes.

Comme dans la plupart des enquêtes par panel, la non-réponse lors du deuxième cycle de l'enquête n'est pas entièrement aléatoire. Elle dépend des modalités de logement du répondant, de son intérêt pour la politique et de quelques variables sociodémographiques (Loosveldt et Carton 1997). Cet abandon sélectif limite la généralisabilité des résultats en ce qui concerne l'évolution de la variable dépendante, mais nos analyses n'indiquent toutefois aucune évolution générale de l'utilisation de la réponse «Ne sait pas». Il n'est pas improbable non plus que la non-réponse sélective lors du deuxième cycle ait un effet sur la grandeur de l'effet d'intervieweur car, comme le montre la première analyse, les caractéristiques du répondant peuvent interagir avec l'effet d'intervieweur. Il est toutefois peu probable que ces interactions influencent les conclusions de fond concernant l'effet d'intervieweur. Étant donné les résultats de la première analyse et les conclusions de l'article de Loosveldt et Carton, on pourrait s'attendre à ce que l'effet d'intervieweur qui se dégage de la deuxième analyse et, par conséquent, l'effet global d'intervieweur, soit quelque peu sous-estimé. Loosveldt et Carton (1997, p. 1021) mentionnent que les répondants les moins instruits sont plus susceptibles de se retirer de l'enquête que ceux dont le niveau de scolarité est plus élevé et la première analyse montre que la variance liée à l'intervieweur est plus forte pour les répondants dont le niveau de scolarité est faible.

Du point de vue méthodologique, les conclusions tiennent compte de l'utilisation de divers modèles pour analyser l'effet d'intervieweur dans le cadre des enquêtes par panel. L'étude présentée ici montre que l'on peut analyser les données à structure compliquée résultant de plans de sondage assez complexes en spécifiant le modèle multiniveaux approprié. Le premier modèle (première analyse) ne s'adapte qu'à un petit nombre de cas. Il n'est pas si courant, ni toujours possible, d'affecter le même intervieweur au même répondant pour divers cycles d'une enquête par panel. Le deuxième modèle (deuxième analyse) est un outil approprié, mais qui exige une capacité de traitement énorme. Le logiciel *MLn* est assez puissant et permet de diminuer la quantité de mémoire requise, au prix d'une légère perte d'information. Toutefois, le deuxième modèle a lui aussi ses limites. Il ne permet pas de modéliser les interactions entre les variables qui caractérisent le répondant et la variance liée à l'intervieweur, comme nous l'avons fait lors de la première analyse, ni celles entre les variables caractérisant le répondant et l'intervieweur. Cependant, l'analyse montre que ce modèle pourrait être un outil fort utile et fort souple. Le modèle à classification croisée convient également si le nombre de mesures augmente. Une enquête par panel comportant trois ou quatre cycles, ou même davantage, où l'on retient certains intervieweurs et où l'on recrute de nouveaux intervieweurs lors de chaque cycle nécessiterait exactement la même analyse. Le modèle multiniveaux permet aussi de traiter les répondants pour lesquels une ou plusieurs mesures manquent,

Le modèle a est le modèle nul: il ne contient pas de variables explicatives, mais la variance de la variable dépendante comprend une composante liée aux mesures, une composante liée au répondant et une composante liée à l'intervieur. La variance liée à l'intervieur est significative. Donc, ce modèle nous donne de nouveau la preuve qu'il existe un effet d'intervieur. Néanmoins, il faut interpréter avec prudence l'importance relative des variances, lorsqu'une classification compte un nombre nettement moins grand d'unités que l'autre (Goldstein 1995, p. 117-118). Il n'est pas tout à fait correct de dire que la variation entre répondants est cinq fois plus grande que la variation entre intervieurs, mais, une fois de plus, nous constatons que la variabilité entre répondants est nettement plus forte qu'entre intervieurs.

Dans le modèle suivant (modèle b), nous avons inclus la variable de temps (ANNEE). De nouveau, l'effet de cette variable n'est pas significatif et son ajout dans le modèle n'améliore pas l'ajustement de celui-ci. Nous pouvons donc conclure que, dans l'ensemble, le nombre de réponses «Ne sait pas» n'évolue pas de façon significative au cours du temps.

Le modèle c est le modèle qui contient les variables caractérisant les répondants. Elles ont toutes un effet significatif et ce modèle est nettement mieux ajusté que les précédents. Pour ce qui est du fond, l'interprétation des paramètres est la même que pour la première analyse. Les femmes recourent à la réponse «Ne sait pas» plus souvent que les hommes et le nombre de ces réponses est plus faible si le niveau de scolarité est élevé. La mesure dans laquelle les répondants suivent les nouvelles politiques dans la presse est également un prédicteur de l'utilisation de la réponse «Ne sait pas». Le nombre de réponses «Ne sait pas» est d'autant plus élevé que l'intérêt pour la politique est faible.

8. CONCLUSION ET DISCUSSION

Les conclusions générales de l'étude portent sur la méthodologie ainsi que sur le fond.

Notre analyse corrobore les résultats d'études antérieures concernant l'utilisation de la réponse «Ne sait pas». Le choix de cette réponse est lié au niveau de scolarité et au sexe du répondant, ainsi qu'à son intérêt pour le sujet. De surcroît, le choix varie vraisemblablement selon l'intervieur. Toutes nos analyses témoignent d'un effet d'intervieur important. Nous ne constatons aucune évolution significative du recours à la réponse «Ne sait pas» d'un cycle à l'autre de l'enquête. L'effet d'intervieur prouve que le choix de la réponse «Ne sait pas» ne résulte pas simplement du processus de réponse du répondant et indique qu'il faut donner une formation aux intervieurs, y compris des instructions sur la façon de poser les questions difficiles et de traiter les réponses «Ne sait pas».

intervieurs demande plus de calculs que l'analyse d'un ensemble de données contenant 10 groupes de 100 mesures classifiées selon 50 répondants et 10 intervieurs. Parfois, il est utile d'omettre certaines observations (combinaisons de mesures à des répondants et à des intervieurs que l'on n'observe pratiquement jamais) pour rendre la séparation plus efficace.

Le logiciel *MLN/MLwIN* offre certaines procédures (grâce aux commandes *XSEArch* et *BXSSEArch*) conçues pour procéder à cette séparation (Rasbash et Woodhouse 1996, 89-93). Nous avons utilisé la commande *BXSSEArch* qui lance une procédure améliorée visant à produire la séparation maximale en réduisant au minimum la suppression des données: au départ, nous avions 4 790 mesures, 3 026 répondants et 275 intervieurs. Après omission des observations indiquées par la commande *BXSSEArch*, il nous restait 4 597 mesures sur 3 026 répondants interviewés par 275 intervieurs. Aucune unité de niveau supérieur (ni répondants ni intervieurs) n'est laissée de côté. La procédure n'a donné que sept sous-ensembles pour lesquels la classification croisée des répondants et des intervieurs a donné, au plus 44 cellules. Appliqué de cette façon, le modèle a convergé suffisamment rapidement.

Les résultats de l'analyse sont présentés au tableau 3.

Tableau 3
Analyse l'ensemble des répondants
(é.-l. entre parenthèses)

Fixe	modèle a	modèle b	modèle c
Niveau des mesures	3,894 (0,136)	3,967 (0,155)	3,864 (0,165)
constante			
année	-0,053 (0,055)		
Niveau des répondants			
sexe	1,808 (0,153)		
scolarité	-1,914 (0,148)		
pressel	1,185 (0,090)		
presses2	1,197 (0,102)		
Aléatoire			
Niveau 2			
Intervieurs			
$\sigma^2_{\text{constante}}$	2,777 (0,373)	2,716 (0,368)	2,844 (0,363)
Répondants			
$\sigma^2_{\text{constante}}$	11,810 (0,635)	11,800 (0,635)	7,017 (0,527)
$\sigma^2_{\text{constante}}$	13,130 (0,475)	13,150 (0,476)	13,460 (0,480)
-2.LL	27717,1	27716,3	27042,1
Δ df *			

Nota: * comparativement au modèle a

Ce tableau rassemble fort au tableau 2, mais présente une différence importante. À la partie aléatoire, nous avons noté le niveau 2, qui englobe les intervieurs et les répondants, pour indiquer clairement que les intervieurs ne représentent pas un troisième niveau dans l'analyse.

Les variables caractéristiques du répondant améliorèrent considérablement l'ajustement du modèle. La diminution de la valeur de $-2 \text{ Log } L$ est importante et nettement significative ($p < 0,001$). Selon l'analyse, les femmes choisissent l'option «Ne sait pas» plus fréquemment que les hommes et les répondants très instruits, moins fréquemment que ceux dont le niveau de scolarité est faible. Suivre les nouvelles politiques dans la presse réduit les chances de répondre «Ne sait pas». Les effets de presse1 et de presse2 sont tous deux significatifs (modèle c). L'inclusion des variables caractérisant les répondants produit aussi une diminution importante de la variance au niveau des répondants.

Nous avons également essayé d'ajuster des pentes aléatoires au niveau des intervieweurs (modèle b). Notre

analyse montre une certaine variation du paramètre associé au niveau de scolarité des répondants. Cette variable indépendante est la seule dont le coefficient est variable au troisième niveau. La valeur de $\sigma^2_{\text{scolarité}}$ n'est pas significative, mais la covariance entre le résidu pour la constante et le résidu pour le niveau de scolarité est importante ($\alpha_{\text{scolarité}/\text{constante}} = -4,099$). La covariance est négative,

recueillant un plus grand nombre de réponses «Ne sait pas», l'écart entre les répondants les moins instruits et les plus instruits sera plus important. Dans le modèle d, la valeur de $\sigma^2_{\text{constante}}$ au niveau de l'intervieweur augmente considérablement comparativement au modèle c. Dans ce modèle, la variance au niveau de l'intervieweur dépend de la valeur de la variable explicative Niveau de scolarité et sera plus importante si cette valeur est nulle. Ceci représente une autre interprétation du modèle d : la variance liée à l'intervieweur est beaucoup plus importante pour les répondants peu instruits que pour les répondants très instruits. Ce

modèle, dont la structure de la variance au niveau 3 est plus complexe, est mieux ajusté que les précédents. L'introduction de la variable ANNEE dans le modèle c ou dans le modèle d, n'a aucun effet significatif non plus. En outre, nos modèles finals ne donnent aucune preuve d'une évolution du nombre de réponses «Ne sait pas» d'un cycle à l'autre. Tous les modèles prouvent qu'il existe un effet d'intervieweur important. Toutefois, la grandeur relative de la variance montre que la variation est plus forte entre répondants qu'entre intervieweurs.

7. DEUXIÈME ANALYSE: ENSEMBLE DES RÉPONDANTS

Dans cette deuxième analyse, la structure hiérarchique est rompue. Les mesures sont encore emboîtées dans les répondants et ceux-ci sont encore emboîtés dans les intervieweurs. Par contre, il n'existe aucune structure hiérarchique globale, puisque l'intervieweur peut changer (et, dans la plupart des cas, change) d'un cycle à l'autre (voir section 3). La variable dépendante demeure le nombre de réponses «Ne sait pas» que donne le répondant i interviewé par l'intervieweur j au moment t (Y_{ijt}). Cependant, le modèle change. L'équation de niveau 1 reste la même:

$$Y_{ijt} = \pi_{0ij} + \pi_{1ij} \text{ANNEE} + e_{ijt}.$$

Ici, nous utilisons π comme notation, puisque le modèle de niveau 1 représente aussi une courbe de croissance. Cependant, cette équation correspond au niveau 1 du modèle à classification croisée (équation (8), section 4.3). En outre, nous continuons d'utiliser l'indice i pour le répondant et l'indice j pour l'intervieweur. Mais il est important de souligner qu'il ne s'agit pas du même modèle que celui utilisé pour la première analyse. Ces indices correspondent aux indices j_1 et j_2 des équations (8) et (9). Il n'existe aucun troisième niveau «réel». Pour ajuster le modèle à classification croisée dans le programme *MLN*, nous devons définir un troisième niveau, mais, conceptuellement, le répondant et l'intervieweur se situent au même niveau dans ce modèle. Nous obtenons ainsi l'équation de niveau 2 suivante:

$$\pi_{0ij} = \pi_0 + \beta_{01} \text{SEX}_i + \beta_{02} \text{SCOLARITE}_i + \beta_{03} \text{PRESSE1}_i + \beta_{04} \text{PRESSE2}_i + r_{0ij} + r_{0j}.$$

La partie propre à l'intervieweur (r_{0j}) est incluse dans le deuxième niveau, si bien qu'il n'y a aucune interaction entre la variance liée à l'intervieweur et les variables qui caractérisent le répondant. Cette différence par rapport à la première analyse est la plus importante.

Le modèle à classification croisée demande une capacité de traitement informatique énorme. Nous avons 3 026 répondants et 275 intervieweurs. Nous devons donc créer 275 variables fictives dont tous les coefficients varient au troisième niveau fictif. À l'heure actuelle, il est impossible d'ajuster un tel modèle. Le chiffré demande beaucoup trop de mémoire (voir Goldstein 1995, p. 118 et Kasbash et Woodhouse 1996, 85-86 pour plus de précision). Il est possible de réduire la mémoire nécessaire et d'améliorer la vitesse d'estimation du modèle en séparant l'ensemble de données en sous-ensembles pour lesquels la classification croisée produit un nombre «relativement» moins élevé de cellules. Ici, nous avons recherché des groupes distincts de mesures, classifiées selon un moins grand nombre de répondants et d'intervieweurs. L'analyse d'un seul groupe de 1 000 mesures classifiées selon 500 répondants et 100

membres du panel qui ont été interviewés deux fois par le même intervieweur, les autres membres du panel et les

Notre variable de temps (*t*), qui est l'année où a eu lieu l'interview, peut prendre la valeur 0 (1992) ou 3 (1995). Nous appliquons un test pour déterminer si π_{ij}^{tj} est un terme significatif. S'il ne l'est pas, le modèle devient un modèle nul ou «modèle naïf» (voir Snijders 1996, 411) selon lequel le nombre de réponses «Ne sait pas» ne varie pas au cours du temps et nous pouvons alors considérer les deux mesures comme des tests répétés d'une même valeur constante. Les coefficients de l'équation de niveau 1 sont propres au répondant et à l'intervieweur.

Au niveau des répondants, nous incluons trois variables, à savoir le sexe, le niveau de scolarité et les deux variables de consultation de la presse. Donc, l'équation de niveau 2 contient quatre variables:

$$\pi_{0ij} = \pi_{0j} + \beta_{01j} \text{SEXE}_i + \beta_{02j} \text{SCOLARITE}_i + \beta_{03j} \text{PRESSE1}_i + \beta_{04j} \text{PRESSE2}_i + r'_{0ij}.$$

Si le paramètre estimé associé à l'année est significatif, nous aurons une équation similaire pour π_{ij}^{tj} .

Au troisième niveau (intervieweurs), nous n'incluons aucune variable supplémentaire, mais nous ajustons une coordonnée à l'origine aléatoire et des pentes aléatoires. Nous obtenons ainsi l'équation de niveau 3 suivante:

Analyse des répondants qui ont été interviewés deux fois par le même intervieweur (é.-l. entre parenthèses)				
Fixe	Modèle a	Modèle b	Modèle c	Modèle d
Niveau des mesures				
constante	4,136 (0,322)	4,028 (0,358)	3,749 (0,442)	3,754 (0,523)
année		0,072 (0,089)		
Niveau des répondants				
sexe		2,393 (0,434)	2,393 (0,434)	2,458 (0,414)
niveau de scolarité		-1,675 (0,425)	-1,675 (0,425)	-1,778 (0,446)
presse1		0,911 (0,263)	0,911 (0,263)	0,887 (0,233)
presse2			1,483 (0,236)	1,426 (0,234)
Aléatoire				
Niveau des intervieweurs				
$\sigma^2_{\text{constante}}$	2,249 (1,040)	2,251 (1,043)	2,666 (0,969)	6,090 (2,109)
$\sigma^2_{\text{scolarité/constante}}$				-4,099 (1,816)
$\sigma^2_{\text{scolarité}}$				1,396 (1,819)
Niveau des répondants				
$\sigma^2_{\text{constante}}$	14,470 (1,714)	14,480 (1,714)	8,939 (1,308)	8,692 (1,332)
Niveau des mesures				
$\sigma^2_{\text{constante}}$	13,320 (0,974)	13,300 (0,974)	13,270 (0,969)	13,250 (0,969)
$\Delta \text{ df}^*$	4519,35	4518,62	4414,52	4395,68

Nota: * comparativement au modèle a

Tableau 2

Analyse des répondants qui ont été interviewés deux fois par le même intervieweur (é.-l. entre parenthèses)

Si nous entrons ces spécifications du modèle dans le programme *MLn*, nous obtenons les résultats qui suivent.

Le modèle a du tableau 2 est le modèle nul. Ce modèle ne contient des variables indépendantes ni au niveau des mesures ni à celui des répondants. Dans le cas de ce modèle, le nombre de réponses «Ne sait pas» est constant. Toutefois, la variance de la variable dépendante comprend une composante liée aux mesures, une composante liée au répondant et une composante liée à l'intervieweur. Toutes les variances sont significatives. Autrement dit, une variation s'est produite d'un cycle à l'autre, certains répondants ont utilisé la réponse «Ne sait pas» plus que d'autres et certains intervieweurs ont obtenu plus de réponses «Ne sait pas» que d'autres.

L'inclusion de la variable Année ne produit pas un modèle mieux ajusté. La diminution de l'écart à la moyenne variable n'est pas significatif non plus (modèle b). Nous pouvons conclure qu'il n'y a eu aucun changement global significatif du nombre de réponses «Ne sait pas». Nous pouvons donc poursuivre avec un modèle sans variable de temps.

Au niveau des répondants, nous considérons trois variables indépendantes: le sexe (0 = homme, 1 = femme), le niveau de scolarité (0 = faible 1 = élevé) et la mesure dans laquelle les répondants suivent les nouvelles politiques (presque 1 = (presque) toujours - 5 = jamais). Les deux premières variables sont constantes pour les deux moments de mesure. La troisième est une covariable qui varie en fonction du temps; qui plus est l'énoncé de la question a été légèrement modifié pour le deuxième cycle. Les deux variantes de la question sur la presse figurent également en annexe. Cette dissimilarité de l'énoncé des questions complicate encore davantage l'établissement du modèle. Le moyen de traiter ce genre de variable consiste à la normaliser (moyenne 0, variance 1) pour chaque moment de mesure, puis de lui attribuer la valeur 0 au moment de la mesure si la question n'a pas été posée. La valeur de référence pour ces deux variables est leur moyenne (voir Snijders 1996, p. 422). Nous obtenons ainsi deux variables, à savoir presse1 pour le premier cycle et presse2, pour le deuxième. La première est nulle pour tous les répondants dans le cas de la deuxième mesure. Nous n'introduisons pas l'âge du répondant dans le modèle, car cette variable serait corrélée trop fortement au moment de la mesure au niveau du cycle d'enquête.

Pour éviter de trop compliquer l'analyse, nous n'introduisons pas de variables caractéristiques de l'intervieweur. Nous supposons simplement qu'il existe un effet d'intervieweur, sans essayer d'expliquer cet effet en fonction des caractéristiques de l'intervieweur.

6. PREMIÈRE ANALYSE: CATÉGORIE 1 DE RÉPONDANTS DU TABLEAU 1

La première analyse ne porte que sur les répondants qui ont été interviewés deux fois par le même intervieweur (c'est-à-dire la catégorie 1 du tableau 1). Cette analyse se fait au moyen d'un modèle «simple» à trois niveaux: mesures emboîtées dans les répondants, eux-mêmes emboîtés dans les intervieweurs. La structure hiérarchique est sans ambiguïté. Ce modèle est semblable à l'exemple du chapitre 8 du traité de Bryk et Raudenbush (1992). Dans cet exemple, les auteurs analysent l'évolution du rendement scolaire des élèves dans les écoles.

Notre variable dépendante est le nombre de réponses «Ne sait pas» pour le répondant i au moment t , interviewé par l'intervieweur j (X_{ij}^{np}). Comme notre analyse ne porte que sur deux mesures, le degré du polymisme ne peut être supérieur à 1. Nous obtenons ainsi l'équation de niveau 1 suivante:

$$X_{ij}^{np} = \pi_{0ij} + \pi_{1ij} \text{ANNEE} + e_{ij}^{np}$$

moyen d'une formulation purement hiérarchique et (conséquemment) d'un logiciel multiniveau type. La méthode consiste à établir les spécifications de l'une des classifications conformément à une classification hiérarchique type, puis à définir une variable fictive pour chaque unité de l'autre classification, à préciser que chaque variable fictive possède un coefficient aléatoire au niveau supérieur et à poser la condition que les ensembles de variances résultants doivent être égaux.

Aux sections 6 et 7, nous servons de ces trois modèles, qui peuvent tous être appliqués au moyen du logiciel de modélisation multiniveaux *MLM/MLwin*. Au préalable, nous examinons plus en détail les variables que nous utilisons dans l'analyse.

5. VARIABLES ANALYSÉES

Durant l'interview de l'enquête post-électorale, l'une des tâches les plus difficiles consistait à évaluer six partis politiques au moyen de diverses échelles à 11 points. Trois échelles ont été présentées au répondant, pour évaluer, respectivement, le catholicisme, le libéralisme économique et le fédéralisme. La question comportait un filtre explicite (Ne sait pas), mais celui-ci n'était pas mentionné sur la carte décrivant les divers choix donnés au répondant. L'énoncé complet de la question figure en annexe. Nous nous attendions à recueillir un nombre considérable de réponses «Ne sait pas», à cause de la complexité de la tâche. Nous nous attendions aussi à ce que le filtre explicite fasse augmenter ce nombre (voir, par exemple Schuman et Presser 1981).

Lors du premier cycle, le nombre moyen de réponses «Ne sait pas» était supérieur à quatre par répondant. Presque 20 % de répondants ont choisi cette option au moins neuf fois sur 18. Si nous considérons uniquement les répondants membres du panel, le nombre moyen est légèrement inférieur (3,8). Ces résultats ne sont pas surprenants, puisque l'on pouvait s'attendre à ce que les «utilisateurs multiples» de la réponse «Ne sait pas» soient sous-représentés lors du deuxième cycle, à cause du manque d'intérêt pour le sujet de l'enquête et (ou) de la difficulté des questions. Au deuxième cycle, le nombre global moyen de réponses «Ne sait pas» était de 3,6 et la moyenne pour les répondants membres du panel se chiffrait à 3,4. Pour les répondants qui ont été interviewés deux fois par le même intervieweur, le nombre moyen est de 3,9 et de 4,2, respectivement. Rien ne permet d'expliquer pourquoi le nombre de réponses «Ne sait pas» enregistré lors du deuxième cycle pour ces répondants est plus élevé que le nombre moyen observé pour l'ensemble des répondants.

Au niveau des mesures, nous utilisons l'année de l'interview comme indicateur du moment de la mesure. Nous avons recodé cette variable de sorte que le temps soit égal à zéro pour le premier cycle et égal à trois pour le deuxième.

intervieweur à l'autre. Pour chaque valeur de β , il existe un résidu lié à l'intervieweur (u_{0j} ou u_{1j}). On peut aussi faire dépendre les coefficients β de variables de plus haut niveau (caractéristiques de l'intervieweur) pour permettre les généralisations sur l'ensemble des intervieweurs. Nous avons une variable de deuxième niveau z . En introduisant (3) par substitution dans (1), nous obtenons le modèle global suivant:

$$Y_{ij} = \beta_0 + \beta_1 x_{1ij} + \gamma_{01} z_{ij} + \gamma_{11} z_{ij} x_{1ij} + u_{1j} x_{1ij} + u_{0j} + e_{ij}. \quad (4)$$

Naturellement, on pourrait inclure un plus grand nombre de variables x et z dans ces équations. Nous supposons que les résidus u_{0j} , u_{1j} et e_{ij} ont une moyenne nulle, étant donné les valeurs des variables explicatives z et x . De surcroît, nous supposons que les résidus de niveau 1 (e_{ij}) sont indépendants. Nous supposons en outre que les résidus de niveau 2 (u_{0j} et u_{1j}) sont indépendants de e_{ij} et qu'ils suivent une loi normale multivariée commune avec matrice de covariance Σ . Ces résidus ne doivent pas être indépendants l'un de l'autre. Habituellement, ils sont corrélés.

4.2 Modèle multiniveaux pour l'analyse longitudinale

Le deuxième modèle que nous utilisons est le modèle multiniveaux longitudinal. Lors de l'analyse de «mesures répétées» au moyen d'un modèle hiérarchique, on considère que les mesures représentent le premier niveau et les unités mesurées, le deuxième. La plupart du temps les unités individuelles sont des personnes, mais il pourrait naturellement s'agir d'autres unités, comme des écoles ou des pays. Dans le cas qui nous occupe, les unités sont des répondants. L'analyse a pour but d'estimer une courbe de croissance d'après les diverses mesures et d'étudier la variation de la courbe selon les caractéristiques individuelles. Chaque valeur observée est subordonnée au moment de la mesure, lequel peut coïncider avec une mesure de temps ou avec l'âge, ainsi qu'aux transformations éventuelles de cette mesure. Habituellement, on suppose que la courbe a une forme polynomiale dont l'équation est la suivante:

$$Y'' = \pi_{0i} + \pi_{1i}t + \pi_{2i}t^2 + \dots + \pi_{ki}t^k + e''_{ii}. \quad (5)$$

Y'' représente la valeur observée pour le répondant i au moment t , t pouvant être le moment où est faite la mesure ou l'âge. Les π_{hi} ($h = 0, \dots, k$) sont les paramètres de trajectoire ou paramètres de croissance pour le sujet i , et k représente le degré du polynôme. Dans un cas simple, k est égal à 1 et la courbe est linéaire. S'il existe m moments de mesure, un polynôme de degré $m - 1$ représentera exactement la courbe. Évidemment, il est plus intéressant d'utiliser un polynôme de faible degré si celui-ci donne une représentation satisfaisante de la courbe. On peut faire l'essai pour voir si le modèle de degré $k + 1$ donne de nettement meilleurs résultats que le modèle de degré k .

On donne aussi aux paramètres de croissance un indice qui représente l'unité (répondant). Le modèle précise que ces paramètres diffèrent d'une unité à l'autre. La deuxième partie du modèle définit ces paramètres:

$$\pi_{0i} = \pi_0 + r_{0i} \quad (6)$$

$$\pi_{0i} = \pi_0 + \beta_{01} x_{1i} + r_{0i}. \quad (7)$$

Le troisième modèle que nous utilisons est le modèle à classification croisée. Les données n'ont pas toute une structure purement hiérarchique. Certaines unités peuvent être classées en fonction de plus d'une variable (voir Goldstein 1995, p. 113-116). Par exemple, des élèves peuvent être classés selon l'école qu'ils fréquentent ou selon le quartier où ils vivent. Dans notre exemple, les mesures sont classées selon le répondant et selon l'intervieweur. Un modèle à classification croisée prend la forme qui suit (les indices j_1 et j_2 désignent les deux structures de classification distinctes):

$$X_{j_1 j_2} = \beta_{0j_2} + \beta_{j_1 j_2} x_{j_1 j_2} + e_{j_1 j_2} \quad (8)$$

$$\beta_{0j_2} = \beta_0 + u_{0j_2} \text{ et } \beta_{j_1 j_2} = \beta_{j_1 j_2} + n_{j_1} + u_{j_1 j_2}. \quad (9)$$

L'équation (9) peut être reformulée de la même façon que l'équation (3). $X_{j_1 j_2}$ représente la valeur observée pour l'unité i , classifiée selon j_1 et j_2 , c'est-à-dire, ici, la valeur de la mesure i faite sur le répondant j_1 , interviewé par l'intervieweur j_2 . Les paramètres associés à la variable indépendante x ont un résidu pour chaque structure de classification. Pour ce modèle, on pose en outre que les résidus obtenus pour des structures de classification différentes (ici les résidus liés au répondant et à l'intervieweur) sont mutuellement indépendants (u_{0j_1} et $u_{j_1 j_2}$ contre u_{0j_2} et $u_{j_1 j_2}$). Raudenbush (1993) examine ce genre de modèles et l'utilisation de l'algorithme EM pour les estimer. Rasbash et Goldstein (1994) et Goldstein (1995, p. 123-124) montrent comment spécifier et estimer ces modèles au

nombre de réponses «Ne sait pas» est la variable dépendante de nos analyses.

3. DESCRIPTION DE LA STRUCTURE DES DONNÉES

Après les élections générales de 1991 en Belgique, on a

réalisé une enquête nationale comptant 4 544 interviews directes réalisées dans les trois régions du pays au cours des premiers mois de 1992. Le plan de sondage est un plan à deux degrés autopondéré (voir, par exemple, Särndal, Swensson et Wretman 1992, p. 141-144). L'échantillon est représentatif de la population de 18 à 74 ans (ISPO/PIOP 1995). Dans le présent article, nous utilisons les données recueillies dans la région flamande, après de 2 691 répondants flamands, interviewés par 163 intervieweurs (Carton, Swyngedouw, Billiet et Beerten 1993). Après les élections de 1995, on a réalisé une enquête post-électorale comparable. Étant donné les contraintes budgétaires, il a fallu sélectionner un échantillon plus petit pour le deuxième cycle. Donc, on a commencé par sélectionner un échantillon à partir du groupe de 2 691 répondants, puis on a sélectionné de nouveaux répondants pour compenser le vieillissement de la cohorte la plus jeune sélectionnée en 1991. En dernière analyse, 2 099 répondants ont été interviewés par 167 intervieweurs. Cet échantillon contient 1 762 répondants appartenant au panel et 337 nouveaux répondants (voir Beerten, Billiet, Carton et Swyngedouw, 1997, pour un rapport technique détaillé sur le plan de sondage). En tout, seuls 55 intervieweurs qui avaient participé au premier cycle ont de nouveau participé à l'enquête. Par conséquent, 112 nouveaux intervieweurs ont participé au deuxième cycle.

Nous obtenons ainsi un ensemble de données correspondant à 3 028 répondants (2 691 + 337) et à 275 intervieweurs (163+112). Pour 1 762 répondants, nous disposons de mesures faites lors de deux cycles, et pour le reste (1 266), d'une seule mesure. On peut représenter la structure de l'ensemble de données dans un tableau compact au tableau 1 (voir aussi Goldstein 1995, p. 114). Chaque x du tableau représente une observation. L'ensemble complet de données contient 4 790 observations ((1 762 \times 2) + 1 266). Chaque catégorie de répondants qui figure dans le tableau représente une occurrence éventuelle dans l'ensemble de données.

Le tableau montre qu'il existe trois groupes de répondants: les répondants membres du panel qui ont été interviewés deux fois par le même intervieweur (catégorie 1), les répondants qui ont participé deux fois à l'enquête, mais qui ont été interviewés par deux intervieweurs différents (catégorie 2) et les répondants qui n'ont été interviewés qu'une seule fois (catégories 3 ou 4). Nos deux analyses se fondent sur ces différentes catégories de répondants. Dans le cadre de la première, nous examinons les répondants dont la situation correspond à celle de la

Le tableau montre aussi que nous pouvons faire la distinction entre les intervieweurs, à savoir ceux qui ont collaboré deux fois à l'enquête (catégories A et B) et ceux qui n'ont collaboré qu'une première fois (catégorie C) ou qu'une deuxième fois (catégorie D). Les intervieweurs de la catégorie B ont collaboré aux deux cycles, mais n'ont jamais interviewé un même répondant deux fois (contrairement à ceux de la catégorie A).

Pour analyser ces données de structure complexe, nous combinons trois modèles distincts que nous présentons à la section qui suit.

4. BRÈVE DESCRIPTION DES DIFFÉRENTS MODÈLES MULTINIVEAUX UTILISÉS POUR LES ANALYSES

4.1 Modèle multiniveaux général

Le premier modèle que nous utilisons est le modèle multiniveaux général, dont la forme est:

$$Y_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + e_{ij} \quad (1)$$

$$\beta_{0j} = \beta_0 + u_{0j} \text{ et } \beta_{1j} = \beta_1 + u_{1j} \quad (2)$$

$$\beta_{0j} = \beta_0 + \gamma_{01}z_{1j} + u_{0j} \text{ et } \beta_{1j} = \beta_1 + \gamma_{11}z_{1j} + u_{1j} \quad (3)$$

L'indice i désigne l'unité de niveau 1 et l'indice j , l'unité de niveau 2. Ici, le niveau 1 correspond au répondant et le niveau 2, à l'intervieweur. Donc, la variable de réponse Y du répondant i , interviewé par l'intervieweur j , dépend de la variable x de ce répondant. Cette relation ressemble à un modèle de régression ordinaire, mais les paramètres sont propres à l'intervieweur. Les coefficients β varient d'un

Tableau 1
Une représentation de l'ensemble de données

N (Intervieweurs)	Cycle				N (Répondants)			
	Catégorie 1 de répondants	Catégorie 2 de répondants	Catégorie 3 de répondants	Catégorie 4 de répondants	Catégorie A d'intervieweurs	Catégorie B d'intervieweurs	Catégorie C d'intervieweurs	Catégorie D d'intervieweurs
47	x	x			x	x		
8			x				x	
108							x	
112								x
	2	1	2	1	374	1388	929	337

catégorie 1. Cette dernière ne compte que 394 répondants qui ont été interviewés deux fois par le même intervieweur. La deuxième analyse porte sur l'ensemble des 3 028 répondants (répondants des catégories 1 à 4 du tableau).

et la façon de l'analyser. À la troisième section, nous décrivons les données en détail, afin de préciser leur structure complexe. La quatrième section traite brièvement des divers modèles que nous combinons. À la cinquième section, nous présentons les variables de l'analyse. Aux sixième et septième sections, nous discutons de l'élaboration des deux modèles que nous utilisons et nous présentons les résultats de l'analyse. Enfin, à la huitième section, nous formulons les conclusions de l'article.

2. RÉPONSE «NE SAIT PAS»

Il est généralement reconnu que l'utilisation d'un filtre «Ne sait pas» ou «Sans opinion» augmente la proportion de répondants qui donnent cette réponse et que l'augmentation proprement dite est fonction de la nature du filtre utilisé (Schuman et Presser 1981, p. 143). Krosnick soutient que répondre «Ne sait pas» est une forme de solution de facilité. Cette situation se présente si le répondant n'est pas motivé à déployer l'effort mental nécessaire pour donner la réponse optimale. La réponse «Sans opinion» est une réponse acceptable, mais elle est le résultat d'un processus cognitif «facile». Choisir la solution de facilité est fonction de la difficulté de la tâche, ainsi que du niveau de connaissances, de compétences et de motivation du répondant. Ce raisonnement théorique corrobore l'observation selon laquelle donner la possibilité de répondre «Ne sait pas» fait augmenter la proportion de répondants qui choisissent cette réponse, particulièrement chez les personnes peu instruites et celles qui accordent peu d'intérêt au problème en question (Krosnick 1991). Selon cette thèse, les caractéristiques des répondants que l'on peut associer à l'aspect cognitif de l'acte consistant à répondre aux questions expliquent en grande partie le choix de la réponse «Ne sait pas». Selon des études antérieures, les caractéristiques pertinentes sont le niveau de scolarité (par exemple, Sudman et Bradburn 1974), l'âge (voir, par exemple, Groves 1989, p. 441-443), le sexe (par exemple, Hox, de Leeuw et Kreft 1991) et l'intérêt pour le sujet (par exemple, Groves 1989, p. 419).

Cependant, répondre à une question comprend non seulement le processus cognitif du répondant, mais aussi un processus de communication (Schwarz et Sudman 1995). Or, l'intervieweur joue un rôle important dans ce second processus. Nombreux sont les articles qui traitent de l'intervieweur en tant que source d'erreur de mesure dans les enquêtes (Groves 1989). Le thème principal de ces études est qu'au lieu d'être des collecteurs «neutres» de données, les intervieweurs influencent les réponses que donnent les personnes qu'ils interrogent. La non-réponse à une question est, elle aussi, sujette à l'effet d'intervieweur, comme l'ont montré il y a déjà longtemps, par exemple, Hanson et Marks (1958), ainsi que Bailar, Bailey et Stevens (1977). Par conséquent, un spécialiste des sciences sociales qui cherche à expliquer les réponses «Ne sait pas» doit inclure dans l'analyse les répondants ainsi que les intervieweurs. Le

généralisations de plus grande portée. répondant et de l'intervieweur permet de procéder à des possibilités de remplacer la variance attribuée au répondant. La mesure au niveau de l'intervieweur et du répondant. La mesure, ainsi que les effets des variables explicatives hiérarchique permet d'estimer la variance liée à l'intervieweur, ainsi que les effets des variables explicatives mesurées au niveau de l'intervieweur et du répondant. La possibilité de remplacer la variance attribuée au répondant et à l'intervieweur par les effets des caractéristiques du répondant et de l'intervieweur permet de procéder à des généralisations de plus grande portée.

L'utilisation du modèle multiniveaux est également fructueuse lors de l'analyse de données longitudinales, c'est-à-dire de «mesures répétées» (voir, par exemple, Goldstein 1995, p. 87-95; Snijders 1996; Yang et Goldstein 1996). Il existe d'autres moyens d'analyser le plan de sondage avec «mesures emboîtées dans les unités», mais l'analyse multiniveaux présente certains avantages manifiestes. Le modèle hiérarchique peut s'appliquer aux plans de sondage non équilibrés, c'est-à-dire ceux où le nombre de mesures n'est pas le même pour toutes les unités observées, et permet d'intégrer assez facilement les covariables qui évoluent. Qui plus est, le modèle permet de tenir compte d'un plus grand nombre de niveaux hiérarchiques. Les unités observées peuvent être regroupées dans une autre unité de niveau supérieur.

Nous analyserons les effets de répondant et d'intervieweur sur le nombre de réponses «Ne sait pas» à une série de questions sur les partis politiques posées lors d'une enquête par panel. Les mesures (cycle 1 et cycle 2) sont emboîtées dans les répondants (plan de sondage longitudinal) et les répondants sont emboîtés dans les intervieweurs. Nos données de panel englobent celles de deux cycles d'enquête. Durant le deuxième cycle, la plupart des répondants n'ont pas été interviewés par le même intervieweur que la première fois, situation qui interrompt le groupement purement hiérarchique pour donner lieu à une structure plus complexe des données. Pour tenir compte de cette structure, il faut concevoir les mesures comme étant groupées dans deux structures de classification distinctes, à savoir les répondants et les intervieweurs. Nous parlerons ici de plan de sondage à classification croisée, car le groupement par niveau n'est pas purement hiérarchique. Dans le présent article, nous partons d'une structure simple de données et du modèle approprié. Puis, le modèle devient plus complexe. Nous réalisons deux analyses. Dans la première, nous nous attachons uniquement aux répondants qui sont interviewés deux fois par le même intervieweur. Ensuite, nous nous intéressons à l'ensemble des répondants, y compris ceux qui n'ont été interviewés qu'une seule fois. Lors de la première analyse, la structure purement hiérarchique reste intacte. Le modèle est un modèle «simple» à trois niveaux: mesures emboîtées dans les répondants qui sont eux-mêmes emboîtés dans les intervieweurs. Dans le cas de la deuxième analyse, nous établissons un modèle à classification croisée. Dans ce modèle, les mesures sont classées selon le répondant et selon l'intervieweur.

À la section suivante, nous décrivons la nature de notre variable dépendante, c'est-à-dire la réponse «Ne sait pas».

Modélisation des effets d'intervieweur dans le cas des enquêtes par panel: Une application

JAN PICKERY et GEERT LOOSVELDT¹

RÉSUMÉ

Le présent article décrit la combinaison de deux applications de modèles multinitiaux. Le modèle multinitiaux convient bien à l'analyse des effets d'intervieweur sur les données d'enquête. On peut aussi l'utiliser pour analyser des données longitudinales, c'est-à-dire des «mesures répétées». Nous analysons un indicateur de la qualité des données que nous appliquons à des données de panel recueillies dans le cadre des enquêtes post-électorales auprès de la population belge. Ces données de panel comprennent les données de deux cycles d'enquête seulement. La plupart des répondants qui ont participé aux deux cycles n'ont pas été interviewés par la même personne les deux fois. Par conséquent, il en résulte une structure complexe où les mesures sont emboîtées dans les répondants et les répondants sont emboîtés dans les intervieweurs, sans toutefois présenter une structure hiérarchique globale: la classification croisée. Nous procédons à deux analyses distinctes interviewées deux fois par la même personne. Puis, nous comparons les résultats de ces deux analyses. Nous concluons que le modèle multinitiaux à classification croisée est un outil très souple et fort utile pour analyser les effets d'intervieweur en cas d'enquête par panel.

MOTS CLÉS: Classification croisée; effets d'intervieweur; enquêtes par panel; modèles multinitiaux; réponse «Ne sait pas».

1. INTRODUCTION

Dans le présent article, nous analysons l'effet des caractéristiques du répondant et de l'intervieweur sur le nombre de réponses «Ne sait pas» recueillies lors de deux cycles de l'enquête par panel menée dans le cadre des enquêtes post-électorales auprès de la population belge. Nous utilisons divers modèles multinitiaux que nous appliquons à un sous-ensemble de l'ensemble de données, d'une part, et à l'ensemble de données complet, d'autre part. Le but principal de l'article est d'illustrer l'utilisation de modèles multinitiaux pour analyser les effets d'intervieweur dans le cas d'une enquête par panel.

Le modèle multinitiaux ou hiérarchique est un outil approprié à l'analyse des données emboîtées, c'est-à-dire à plusieurs degrés, comme les élèves groupés dans des classes ou les patients groupés dans des hôpitaux. Un modèle multinitiaux peut inclure les variables qui correspondent aux divers niveaux de groupement, mais aussi tenir compte de la variabilité associée à chaque niveau. La qualité type de ces modèles tient non pas à la forme fonctionnelle établissant le lien entre les variables des divers niveaux, mais plutôt au traitement plus complexe de la structure de l'erreur (Diffré et Fortin 1994, p. 334). Par exemple, dans le cas de la recherche en éducation, un modèle multinitiaux permet de rendre compte de la variation entre écoles et de la variation entre élèves. De surcroît, on s'efforce de remplacer dans le modèle cette variabilité attribuable aux deux niveaux par des variables liées à chaque niveau. Ces modèles sont décrits dans divers manuels, comme Bryk et Raudenbush (1992), Goldstein

(1995), Kreft et de Leeuw (1998) et Snijders et Bosker (1999).

Les modèles multinitiaux ou hiérarchiques sont aussi ceux qui conviennent le mieux à l'analyse des effets d'intervieweur sur les données d'enquête (Hox 1994). Le modèle hiérarchique est l'outil le plus approprié en cas de plan de sondage avec «répondants nichés dans les intervieweurs». L'application d'autres méthodes statistiques exige que les caractéristiques des intervieweurs et des répondants soit mutuellement indépendantes, condition qui, la plupart du temps, n'est pas satisfaite à cause de la structure hiérarchique des données. Dans le cas d'un modèle multinitiaux, aussi bien les coefficients de régression que les composantes de la variance sont subordonnées aux variables explicatives du modèle, propriété utile si l'orthogonalisation des variables caractérisant les intervieweurs et les répondants est incomplète (Hox 1994, p. 307). Si les répondants ne sont pas affectés au hasard aux intervieweurs, les caractéristiques des répondants et des intervieweurs peuvent éventuellement se confondre puisque les répondants d'une région particulière seront vraisemblablement interviewés par des intervieweurs de cette région. Le cas échéant, si les variables pertinentes caractérisant les répondants sont introduites dans un modèle multinitiaux, l'égalisation des variables relatives aux intervieweurs se fait par des moyens statistiques. Par conséquent, les hypothèses qui sous-tendent l'analyse des effets d'intervieweur au moyen d'un modèle multinitiaux sont plus réalistes que celles fournies dans le cas d'un analyse de la variance ou d'un analyse de la covariance. De surcroît, le modèle

¹ Jan Pickery et Geert Loosveldt, Department of Sociology, University of Leuven, E. Van Evenstraat 2B, 3000 Leuven, Belgium.
courriel : jan.pickery@soc.kuleuven.ac.be; geert.loosveldt@soc.kuleuven.ac.be.

dit que la droite de régression ne passe pas l'origine. Dans le cas de cette population, l'estimateur par régression t^* est plus efficace que l'estimateur par quotient d^* . Nous constatons aussi que la taille initiale optimale d'échantillon, n_o , est plus grande pour t^* que pour l'estimateur d^* . Pour la taille optimale d'échantillon de deuxième phase, n_o , nous observons l'inverse, parce qu'on peut obtenir un estimateur par régression plus précis avec un plus petit échantillon de deuxième phase, si bien qu'il est possible d'allouer davantage à l'échantillon de première phase. Enfin, le taux d'échantillonnage inverse optimal k_o est pratiquement le même pour les deux estimateurs. Si la droite de régression passe par l'origine, l'avantage de t^* sur d^* disparaît, comme le prédit et le confirme une autre comparaison empirique que nous ne présentons pas ici.

Tableau 2
Efficacité relative de d^* et t^* par rapport à y^* ($C^* = 200, c = 0,5, c_1 = 1$)

c'	c_2	k_{oHH}	n_{oHH}	k_o	n_o	n_o'	Efficacité
0,1	2	1,58	127	1,46	92	514	1,85
0,1	4	2,23	115	2,06	85	477	1,91
0,3	2	1,58	127	1,46	78	250	1,23
0,3	4	2,23	115	2,06	73	234	1,32
0,1	2	1,58	127	1,47	89	563	2,11
0,1	4	2,23	115	2,08	83	523	2,19
0,3	2	1,58	127	1,47	74	269	1,34
0,3	4	2,23	115	2,08	70	253	1,45

5. CONCLUSIONS

Nous proposons des estimateurs par quotient ou par régression fondés sur une méthode d'échantillonnage à deux phases pour tenir compte de la non-réponse au sujet de la variable principale quand on ne connaît pas la moyenne de population de la variable auxiliaire. Nous éliminons le biais éventuellement important dû à la non-réponse par

Nous remercions les examinateurs et les rédacteurs en chef adjoints de leurs commentaires qui nous ont aidés à améliorer notre article.

REMERCIEMENTS

sous-échantillonnage des non-répondants, conformément à la méthode de Hansen et Hurwitz (1946). Nous calculons les tailles optimales d'échantillon pour un ensemble donné de coûts unitaires, puis nous comparons théoriquement et empiriquement la performance de nos estimateurs à celle de l'estimateur de Hansen et Hurwitz. S'il existe une relation linéaire prononcée entre la variable principale et la variable auxiliaire et que l'on peut recueillir à faible coût les données sur la variable auxiliaire auprès d'un échantillon de grande taille, nos estimateurs donnent de nettement meilleurs résultats que l'estimateur de Hansen et Hurwitz. Notre méthode pourrait être utile s'il existe un biais important dû à la non-réponse que l'on ne peut corriger par rajustement de la pondération ni par imputation.

BIBLIOGRAPHIE

COCHRAN, W.G. (1977). *Sampling Techniques*. 3^e édition. New York: John Wiley & Sons.
HANSEN, M. H., et HURWITZ, W. N. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association*, 41, 517-529.
KALTON, G., et KASPRZYK, D. (1986). Le traitement des données d'enquête manquantes. *Techniques d'enquête*, 12, 1-17.
OH, H.L., et SCHEUREN, F.J. (1983). Chapter 13. Weighting adjustment for unit non-response. Dans *Incomplete Data in Sample Surveys*. (I. Olkin, W.G. Madow, et D.B. Rubin, Eds.). Theory and Bibliographies, 2, 143-184. New York: Academic Press.
RANCOURT, E., LEE, H., et SÄRNDAAL, C.-E. (1994). Corrections du biais pour des estimations d'enquête tirées de données comprenant des valeurs imputées par quotient par suite d'une non-réponse selon un mécanisme avec confusion. *Techniques d'enquête*, 20, 143-153.

où y^* est défini comme dans (2.2).

$$V(y^*) - V(d^*) = \left(\frac{1}{2} - \frac{n}{N} \right) (2RS_{xy} - R^2S_x^2)$$

$$+ \frac{n}{W^2(k-1)} (2RS_{2xy} - R^2S_{2x}^2). \quad (4.2)$$

La différence est positive (autrement dit, d^* est plus efficace que y^*) si $R < 2\beta_2$ et $R < 2\beta_2$, où $\beta_2 = S_{2xy}/S_{2x}^2$.

Par ailleurs, nous avons

$$V(y^*) - V(t^*) = \left(\frac{1}{2} - \frac{n}{N} \right) \left(\frac{S_{xy}^2}{S_x^2} \right)$$

$$+ \frac{n}{W^2(k-1)} \beta S_{2x}^2 (2\beta_2 - \beta). \quad (4.3)$$

où

$$h = \frac{S_x^2}{S_y^2}, \theta_1 = \frac{n_o}{n_o}, \theta_2 = \frac{n_o}{n_o}, \tilde{O}_{HHy} = \frac{n_o}{n_o}, W^2(k_{HHH} - 1) S_{2y}^2, \\ \tilde{O}_n = \frac{S_{2y}^2}{W^2(k_o - 1) S_{2n}^2}, n = x, y,$$

$$2p - Rh > \frac{1}{1 - \theta_1} \times \left\{ \frac{1}{\beta h} (1 - \theta_2 + \tilde{O}_y - \theta_2 \tilde{O}_{HHy}) - h \tilde{O}_x (2\beta_2 - R) \right\} \quad (4.6)$$

Si nous comparons cette expression à celle de $V(d^*)$ avec les valeurs optimales de k, n et n' données par (3.3), alors la condition voulant que d^* soit plus précise que y^* s'écrit

$$V(y^*) = \left(\frac{1}{2} - \frac{n}{N} \right) \left(\frac{S_{xy}^2}{S_x^2} \right) + \frac{n}{W^2(k_{HHH} - 1) S_{2y}^2} \quad (4.5)$$

et où p représente le coefficient de corrélation de x et y . Nous pouvons procéder à une comparaison similaire entre y^* et t^* . Autrement dit, t^* est plus efficace que y^* si

$$2p - \beta h > \frac{1}{1 - \theta_1} \times \left\{ \frac{1}{\beta h} (1 - \theta_2 + \tilde{O}_y - \theta_2 \tilde{O}_{HHy}) - h \tilde{O}_x (2\beta_2 - \beta) \right\}. \quad (4.7)$$

4.3 Comparaison empirique des estimateurs proposés

Nous nous servons d'une population générée artificiellement pour comparer l'efficacité relative des estimateurs d^* et t^* par rapport à y^* . Les paramètres de la population sont les suivants:

$$R = 1,92, \beta = 1,52, \rho = 0,85, R_2 = 1,88, \beta_2 = 1,47, \\ p_2 = 0,83, N = 1000, N_2 = 302, S_x^2 = 766,54,$$

$$S_y^2 = 2426,82, S_{xy} = 1164,08, S_{2x}^2 = 433,63, \\ S_{2y}^2 = 1350,05 \text{ et } S_{2xy} = 638,32.$$

Les efficacités relatives de d^* et t^* sont présentées au tableau 2. Notons que R diffère nettement de β , autrement

Alors, la variance de l'estimateur de Hansen-Hurwitz devient

$$n_{oHH} = \frac{c + c_1 W_1 + c_2 W_2 / k_{oHH}}{C^*} \text{ et} \quad (4.4)$$

$$k_{oHH} = \sqrt{\frac{c_2 (S_y^2 - W_2 S_{2y}^2)}{S_{2y}^2 (c + c_1 W_1)}}.$$

semblable à celui donné par (3.2) selon la même technique (c'est-à-dire, le multiplicateur de Lagrange) que celle utilisée à la section 3, comme suit:

$$C^* = \left(c + c_1 W_1 + \frac{c_2 W_2}{k} \right) n,$$

Nous allons maintenant comparer les estimateurs proposés à l'estimateur de Hansen-Hurwitz (y^*) en nous servant de la fonction de coût donnée à la section 3. Pour l'estimateur y^* , si nous sélectionnons un échantillon aléatoire simple (sans recourir à l'échantillonnage à deux phases) pour y , nous pouvons calculer la taille optimale d'échantillon pour un coût prévu.

Considérons pour d^* la fonction de coût donnée par

$$C = c'n' + cn + c_1n_1 + c_2m \quad (3.1)$$

où les c représentent les coûts unitaires définis comme suit:

c' : coût unitaire associé à l'échantillon de première phase, a_1' ;

c : coût unitaire associé au premier effort de collecte de renseignements sur y auprès de l'échantillon de deuxième phase, a_2 ;

c_1 : coût unitaire du traitement des renseignements sur y fournis par les répondants lors du premier effort de collecte auprès de a_1 ;

c_2 : coût unitaire associé au sous-échantillon a_{2m} de a_2 .

Puisqu'on ne connaît pas la valeur de n_1 tant qu'on n'a pas fait la première collecte de données, on se sert du coût prévu pour la minimisation. Le coût prévu est donné par

$$E(C) = C^* = c'n' + \left(c + c_1W_1 + \frac{k}{c_2W_2} \right) n. \quad (3.2)$$

Nous nous servons du multiplicateur de Lagrange pour recalculer les valeurs optimales de k, n et n' qui réduisent au minimum la variance de d^* pour un coût prévu fixe C^* . Les valeurs optimales ainsi obtenues sont:

$$k_o = \sqrt{\frac{c_2^2(S_2^2 - W_2S_{2r}^2)}{c_2^2(S_2^2 - W_2S_{2r}^2) + c_1^2W_1^2}},$$

$$n_o = \frac{C^*\sqrt{A}}{D\sqrt{G}} \text{ et } n_o' = \frac{D\sqrt{G}}{C^*\sqrt{S_2^2 - S_{2r}^2}}$$

(3.3)

où

$$A = S_2^2 + W_2(k_o - 1)S_{2r}^2,$$

$$G = c + c_1W_1 + \frac{k_o}{c_2W_2} \text{ et}$$

$$D = \sqrt{(S_2^2 - S_{2r}^2)c'} + \sqrt{AG}.$$

Si nous supposons que $\gamma = c_2/(c + c_1W_1)$, $\delta = S_{2r}^2/S_2^2$ et $\xi = S_{2r}^2/S_2^2$, alors nous obtenons

$$V(d^*) = \left(\frac{1}{1} - \frac{1}{N} \right) \left(S_2^2 + \frac{S_{2r}^2}{W_2(k - 1)} \right) \frac{n}{S_2^2} \quad (4.1)$$

donnée par

La variance de l'estimateur de Hansen-Hurwitz est

4.1 Sans tenir compte du coût

À la présente section, nous comparons théoriquement la performance des estimateurs proposés à celle des estimateurs de Hansen et Hurwitz (1946), d'abord sans tenir compte du coût, puis en tenant compte de celui-ci.

4. COMPARAISON DES ESTIMATEURS

C^*	c'	c	c_1	c_2	ξ	W_2	γ	k_o	G	n_o	n_o'
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	382
200	0,1	0,5	1	2	1	2	0,3	1,67			

$$S_y^2 = \frac{n-1}{1} \left\{ \sum_{a_1} y_i'^2 + k \sum_{a_2m} y_i'^2 - n\bar{y}^2 + w_2(k-1)S_z^2 \right\} \quad (2.6)$$

$$S_y^2 = \frac{n-1}{1} \left\{ \sum_{a_1} (y_i' - r^* x_i')^2 + k \sum_{a_2m} (y_i' - r^* x_i')^2 \right\} \text{ et } (2.7)$$

$$S_{zr}^2 = \frac{n-1}{1} \sum_{a_2m} (y_i' - r^* x_i')^2. \quad (2.8)$$

Il convient de souligner que S_y^2 est un estimateur non biaisé de S_y^2 . Il semble naturel d'utiliser S_y^2 pour estimer S_y^2 puisque l'expression obtenue d'après S_y^2 en remplaçant r^* par R est un estimateur convergent de S_y^2 . Nous pouvons nous servir du même argument pour justifier l'emploi de S_{zr}^2 .

Nous obtenons un autre estimateur de $V(d^*)$ en remplaçant S_y^2 par S_{zr}^2 , c'est-à-dire :

$$S_z^2 = S_y^2 + r^{*2} S_x^2 - 2r^* S_{xy}^2 \text{ et } (2.9)$$

$$S_{zr}^2 = S_z^2 + r^{*2} S_x^2 - 2r^* S_{zxy}^2, \quad (2.10)$$

respectivement, dans (2.5), où

$$S_x^2 = \frac{n-1}{1} \sum_{a_1} (x_i' - \bar{x}')^2,$$

$$S_{zmy}^2 = \frac{n-1}{1} \sum_{a_2m} (y_i' - \bar{y}_{2m})^2,$$

$$S_{zx}^2 = \frac{n-1}{1} \sum_{a_2} (x_i' - \bar{x}_2')^2,$$

$$S_{zmy}^2 = \frac{n-1}{1} \left(\sum_{a_2m} x_i' y_i' - m \bar{x}_{2m} \bar{y}_{2m} \right)$$

et où s_{xy}^* a la même forme que dans (2.9). La variance de cet estimateur de recouvrement sera vraisemblablement plus faible que celle de l'estimateur (2.5), puisque les estimateurs $s_{x_2}^2$ et $s_{z_2}^2$ sont fondés sur des échantillons plus grands et sont par conséquent plus précis.

2.3 Estimateur par régression avec échantillonnage à deux phases

Nous définissons l'estimateur par régression comme suit :

$$t^* = y^* + \beta^* (x' - \bar{x}') \quad (2.8)$$

où β^* est un estimateur de $\beta = S_{xy}/S_x^2$. Il pourrait exister plusieurs solutions pour β^* , mais un choix naturel semble être $\beta^* = s_{xy}^*/s_x^{*2}$, où

3. CHOIX DES FRACTIONS D'ÉCHANTILLONNAGE

Nous allons maintenant déterminer les valeurs optimales de k , n et n' qui réduisent au minimum la variance des estimateurs proposés pour un coût particulier ou qui réduisent au minimum le coût pour une variance particulière.

$$S_{z1}^2 = s_{zmy}^2 + \beta^{*2} s_{zx}^2 - 2\beta^* s_{zmy}^2 \quad (2.13)$$

$$S_z^2 = S_y^2 + \beta^{*2} s_x^2 - 2\beta^* s_{xy}^2 \text{ et } (2.12)$$

Comme pour (2.7), nous obtenons un estimateur un peu meilleur de $V(t^*)$ en utilisant :

$$y_i' = y_i^* - \beta^* (x_i' - \bar{x}'). \quad (2.12)$$

$$S_{z1}^2 = \frac{n-1}{1} \sum_{a_2m} (y_i' - y_i^*)^2, \text{ et } (2.11)$$

$$S_z^2 = \frac{n-1}{1} \left\{ \sum_{a_1} (y_i' - y_i^*)^2 + k \sum_{a_2m} (y_i' - y_i^*)^2 \right\},$$

où

$$V(t^*) = \left(\frac{1}{1} - \frac{N}{1} \right) S_y^2 + \left(\frac{n'}{1} - \frac{n}{1} \right) S_z^2 + \frac{n}{w_2(k-1)} S_{z1}^2 \quad (2.11)$$

Pour estimer $V(t^*)$, nous pouvons appliquer la formule suivante :

$$V(t^*) = \left(\frac{1}{1} - \frac{N}{1} \right) S_y^2 + \left(\frac{n'}{1} - \frac{n}{1} \right) S_z^2 + \frac{n}{w_2(k-1)} S_{z1}^2 \quad (2.10)$$

Il est facile de montrer que s_{xy}^* et s_{z2}^2 sont des estimateurs non biaisés de S_{xy}^2 et S_{z2}^2 , respectivement. La variance approximative de t^* est donnée par

$$s_{x_2}^2 = \frac{n-1}{1} \left(\sum_{a_1} x_i'^2 + k \sum_{a_2m} x_i'^2 - n\bar{x}^2 \right). \quad (2.9)$$

$$s_{xy}^2 = \frac{n-1}{1} \left(\sum_{a_1} x_i' y_i' + k \sum_{a_2m} x_i' y_i' - n\bar{x}\bar{y} \right) \text{ et } (2.10)$$

phase de grande taille n' à partir de N unités de la population par échantillonnage aléatoire simple sans remise (FASSR). Puis, nous sélectionnons par EASSR un échantillon de deuxième phase de plus petite taille n à partir de n' et nous mesurons la caractéristique y sur cet échantillon. L'estimateur par quotient de la moyenne de y est $\bar{y}_p = (\bar{y}/\bar{x})\bar{x}$, où \bar{x} est la moyenne d'échantillon calculée pour n' unités, et où \bar{y} et \bar{x} sont obtenues d'après l'échantillon de deuxième phase s'il n'y a pas de non-réponse dans cet échantillon. Cependant, en cas de non-réponse dans l'échantillon de deuxième phase, nous pouvons utiliser un estimateur fondé uniquement sur les répondants et reprendre un sous-échantillon de non-répondants et reprendre contact avec eux. La première option est beaucoup moins coûteuse que la seconde, parce que recueillir les renseignements manquant auprès des non-répondants en les contactant de nouveau nécessite habituellement beaucoup plus d'efforts et de dépenses. Cependant, il se pourrait fort bien qu'en ce qui concerne la caractéristique étudiée, les non-répondants diffèrent des répondants au point que les résultats soient sérieusement biaisés. Le cas échéant, le sous-échantillonnage des non-répondants pourrait être souhaitable. Par conséquent, nous pourrions l'idée du sous-échantillonnage de Hansen et Hurwitz dans le cas d'un échantillonnage à deux phases. Fondamentalement, les estimateurs que nous proposons ici constituent une version à échantillonnage à deux phases des estimateurs proposés par Cochran (1977, p. 374), c'est-à-dire des estimateurs de Y par quotient ou par régression avec échantillonnage à deux phases, corrigés pour la non-réponse par la méthode de Hansen et Hurwitz (1946).

Supposons que les n' unités fournissent toutes des renseignements sur la variable auxiliaire x à la première étape d'échantillonnage. Mais posons que n_1 unités fournissent des renseignements sur y et que n_2 unités refusent de répondre lors de la deuxième étape. À partir des n_2 non-répondants, nous sélectionnons un échantillon aléatoire simple sans remise de m unités au taux inverse d'échantillonnage k , où $m = n_2/k$, $k > 1$. Cette fois-ci, les m unités répondent toutes. Ces conditions pourraient s'appliquer à une enquête-ménages où l'on se sert de la taille du ménage comme variable auxiliaire pour l'estimation, disons, des dépenses familiales. On pourrait obtenir des renseignements complets sur la taille de la famille durant l'établissement de la liste des ménages, mais faire face à une non-réponse en ce qui concerne les dépenses du ménage.

Dans l'exposé qui suit, nous supposons que l'ensemble de la population (représenté par A) est divisé en deux strates: une strate (représentée par A_1) de N_1 unités qui répondent lors de la première visite à la deuxième étape et une strate (représentée par A_2) de N_2 unités qui ne répondent pas lors de la première visite à la deuxième étape d'échantillonnage, mais qui répondent lors de la deuxième visite. Représentons les échantillons de premier et de deuxième phase par a et a' , respectivement, et posons que $a_1 = a \cap A_1$ et $a_2 = a \cap A_2$. Le sous-échantillon de a_2 sera

$$v(d^*) = \left(\frac{1}{1} - \frac{N}{1} \right) S_2^y + \left(\frac{n'}{1} - \frac{1}{1} \right) S_2^x + \left(\frac{n}{w_2(k-1)} - \frac{1}{S_2^2} \right) S_2^y \quad (2.5)$$

Nous pouvons estimer la variance approximative de d^*

$$S_2^2 = S_2^y + R S_2^x - 2 R S_{2xy}, \quad (2.4)$$

$$S_2^2 = S_2^y + R S_2^x - 2 R S_{2xy}, \quad (2.4)$$

où

$$v(d^*) = \left(\frac{1}{1} - \frac{N}{1} \right) S_2^y + \left(\frac{n'}{1} - \frac{1}{1} \right) S_2^x + \left(\frac{n}{w_2(k-1)} - \frac{1}{S_2^2} \right) S_2^y \quad (2.3)$$

linéarisation par série de Taylor, est donnée par

Pour un grand échantillon, l'approximation de premier ordre de la variance de d^* , calculée selon la méthode de

exemple, $x' = (1/n') \sum_{i \in a'} x_i$, de première phase a' sont marqués d'un signe prime (par $\bar{u}_1 = (1/n') \sum_{i \in a_1} u_i$, et celles calculées d'après l'échantillon sont marquées de l'indice «1», (par exemple, $\bar{u}_{2m} = (1/m) \sum_{i \in a_{2m}} u_i$; celles calculées d'après a_1 (par exemple, $\bar{u}_{2m} = (1/m) \sum_{i \in a_{2m}} u_i$), sont marquées de l'indice «2m», $\bar{u}_j = N_j/N$ et $w_j = n_j/n$, $j = 1$ ou 2. Les statistiques d'échantillon calculées d'après a_{2m} sont marquées de l'indice «2m»,

$$\bar{u}' = w_1 \bar{u}_1 + w_2 \bar{u}_{2m}, \quad u = x, y. \quad (2.2)$$

de Hansen-Hurwitz qui sont donnés par

$$d^* = \frac{\bar{y}}{\bar{x}} \bar{x} = r^* \bar{x} \quad (2.1)$$

Nous définissons l'estimateur par quotient avec échantillonnage à deux phases comme suit:

2.2 Estimateur par quotient avec échantillonnage à deux phases

En règle générale, les paramètres de population sont représentés par des lettres majuscules, sauf les lettres grecques, et les statistiques d'échantillon, par les lettres minuscules correspondantes.

représenté par a_{2m} . La somme sur l'ensemble des unités est un ensemble s que nous représentons par \sum_s .

Echantillonnage à deux phases pour l'estimation par quotient ou par régression avec sous-échantillonnage des non-répondants

FABIAN C. OKAFOR et HYUNSHIK LEE¹

RÉSUMÉ

Cochran (1977, p. 374) a proposé certains estimateurs par quotient ou par régression de la moyenne de population fondés sur la méthode de Hansen et Hurwitz (1946) consistant à sous-échantillonner les non-répondants en supposant que l'on connaît la moyenne de population de la variable auxiliaire. Le présent article décrit certains estimateurs par quotient ou par régression axés sur un échantillonnage double (à deux phases) applicables aux cas où l'on ne connaît pas la moyenne de population de la variable auxiliaire. On y compare aussi la performance de ces estimateurs à celle de l'estimateur proposé par Hansen et Hurwitz (1946).

MOTS CLÉS : Estimateur de Hansen et Hurwitz; coût d'enquête; fraction optimale d'échantillonnage.

1. INTRODUCTION

Très souvent, lors d'enquêtes visant des personnes, on ne réussit pas à recueillir les renseignements auprès de toutes les unités d'échantillonnage, même après plusieurs rappels. Or, une estimation calculée d'après des données incomplètes peut être trompeuse, surtout si les caractéristiques des répondants diffèrent de celles des non-répondants, auquel cas l'estimation risque d'être biaisée. Hansen et Hurwitz (1946) ont proposé une méthode de correction pour la non-réponse afin de résoudre le problème du biais. Cette méthode consiste à sélectionner un sous-échantillon de non-répondants afin d'obtenir une estimation pour la sous-population que ces derniers représentaient.

En s'appuyant sur la méthode de Hansen et Hurwitz (1946), Cochran (1977) a proposé les estimateurs par quotient ou par régression de la moyenne de population de la variable étudiée pour lesquels les renseignements sur la variable étudiée pour les unités d'échantillonnage auxiliaire proviennent de toutes les unités d'échantillonnage, alors que certaines unités n'ont pas toutes fourni les renseignements sur la variable étudiée. En outre, on connaît la moyenne de population de la variable auxiliaire. Ici, nous supposons que l'on ne connaît pas la moyenne de population de la variable auxiliaire. Par conséquent, nous utilisons la méthode d'échantillonnage à deux phases pour estimer d'abord la moyenne de la variable auxiliaire, puis la moyenne de la variable étudiée selon une méthode similaire à celle de Cochran (1977).

En pratique, on compense souvent la non-réponse par rajustement de la pondération (Oh et Scheuren 1983) ou par imputation (Kaltou et Karsprzyk 1986). Les méthodes appliquées pour rajuster la pondération ou pour procéder à l'imputation visent à éliminer le biais dû à la non-réponse. Cependant, cette méthode se fonde sur des hypothèses insoutenables quant au mécanisme de réponse. Si le mécanisme hypothétique est erroné, les estimations résultantes

risquent d'être fortement biaisées. De surcroît, il est difficile d'éliminer entièrement le biais s'il y a confusion de la non-réponse, en ce sens que la probabilité de réponse dépend de la variable étudiée. Ramanouj, Lee et Samdal (1994) ont réussi à corriger partiellement cette situation. La méthode de sous-échantillonnage de Hansen et Hurwitz ne présente pas ce défaut, mais elle est plus coûteuse à cause du travail supplémentaire qu'exige le sous-échantillonnage des non-répondants. Néanmoins, si le biais est important, la méthode offre un moyen viable de résoudre le problème sans recourir à la réponse totale qui pourrait coûter fort cher.

À la section suivante, nous examinons les estimateurs par quotient ou par régression axés sur l'échantillonnage à deux phases. En général, on recourt à l'échantillonnage à deux phases lorsque l'on veut servir de données auxiliaires pour améliorer la précision d'une estimation, mais que l'on ne connaît pas la loi de distribution de la population pour les variables auxiliaires. Nous nous servons de l'échantillon de la première phase pour estimer la distribution de population de la variable auxiliaire et de l'échantillon de la deuxième phase pour obtenir les renseignements nécessaires sur la variable étudiée. Nous calculons la fraction optimale d'échantillonnage pour les divers estimateurs, pour un coût prédéterminé. Enfin, nous comparons théoriquement et empiriquement les estimateurs proposés à l'estimateur de Hansen et Hurwitz.

2. ESTIMATEURS PAR QUOTIENT OU PAR RÉGRESSION AVEC ÉCHANTILLONNAGE À DEUX PHASES

2.1 Renseignements généraux

Pour estimer la moyenne de population \bar{X} de la variable auxiliaire, nous sélectionnons un échantillon de première

¹ Fabian C. Okafor, Dept. of Statistics, University of Nigeria, Nsukka, Nigeria; Hyunshik Lee, anciennement Statistique Canada, maintenant Westat, 1650 Research Boulevard, Rockville, Maryland, 20850, U.S.A.

$$\frac{\partial^2 h(\eta)}{\partial^2 \eta} = -m\psi'(\eta) \leq 0$$

puisque $m > 0$ et $\psi'(\eta)$ est positif sur $(0, \infty)$ (Temme, 1994, 54-55).

BIBLIOGRAPHIE

BATTESE, G.E., HARTER, R.M., et FULLER, W.A. (1988). An error components model for prediction of county crop area using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

BEST, N., COWLES, M.K., et VINES, K. (1996). CODA, *Convergence Diagnosis and Output Analysis Software for Gibbs Sampling Output*, Version 0.30. MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR.

GELFAND, A.E. (1995). Model determination using sampling-based methods. Dans *Markov Chain Monte Carlo in Practice* (W.R. Gilks, S. Richardson, et D.J. Spiegelhalter, Eds.), 145-161. London: Chapman and Hall.

GELFAND, A.E., et SMITH, A.F.M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.

GELFAND, A.E., et SMITH, A.F.M. (1991). Gibbs sampling for marginal posterior expectations. *Communications In Statistics - Theory and Methods*, 20, 1747-1766.

GHOSH, M., et RAO, J.N.K. (1994). Small area estimation: An appraisal (avec discussion). *Statistical Science*, 9, 55-93.

GILKS, W.R., BEST, N.G., et TAN, K.K.C. (1995). Adaptive rejection Metropolis sampling within Gibbs sampling. *Journal of Applied Statistics*, 44, 455-472.

YOU, Y., et RAO, J.N.K. (1999). Pseudo hierarchical Bayes small area estimation using sampling weights. *Recueil de la section des méthodes d'enquête, Société Statistique du Canada*, 117-122.

WILEY.

TEMME, N.M. (1994). *Special Functions: An Introduction to the Classical Functions of Mathematical Physics*. New York: John Robinson Way, Cambridge CB2 2SR.

SPIEGELHALTER, D., THOMAS, A., BEST, N., et GILKS, W. (1996). *BUGS 0.5, Bayesian Inference Using Gibbs Sampling Manual*. MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR.

RAO, J.N.K. (1999). Quelques progrès récents concernant l'estimation régionale fondée sur un modèle. *Techniques d'enquête*, 25, 199-212.

PRASAD, N.G.N., et RAO, J.N.K. (1990). The estimation of mean squared error of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.

MOURA, F., et HOLT, D. (1999). Production d'estimations régionales à partir de modèles multiniiveau. *Techniques d'enquête*, 25, 81-89.

KLEFFE, J., et RAO, J.N.K. (1992). Estimation of mean square error of empirical best linear unbiased predictors under a random error variance linear model. *Journal of Multivariate Analysis*, 43, 1-15.

HOLT, D., et MOURA, F. (1993). Small area estimation using multi-level models. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 21-30.

intégrer les poids de plan d'enquête, en conformité avec You et Rao (1999).

REMERCIEMENTS

Nous tenons à remercier deux examinateurs et le rédacteur de leurs remarques et suggestions utiles. Nous tenons également à remercier le professeur F. Moura de l'Université fédérale de Rio de Janeiro (Brésil) d'avoir fourni l'ensemble de données utilisé à la section 3. Nos travaux ont été appuyés en partie par une subvention du Conseil de recherches en sciences naturelles et en génie du Canada.

ANNEXE

A1:

L'estimateur Rao-Blackwellisé de la variance a posteriori de β_j est donné par:

$$V(\beta_j) = \frac{1}{G} \sum_{k=1}^G V(\beta_j | Y, y^{(k)}, \Omega^{(k)}, \tau^{(k)}) + \frac{1}{G} \sum_{k=1}^G [E(\beta_j | Y, y^{(k)}, \Omega^{(k)}, \tau^{(k)})]^2 - \left[\frac{1}{G} \sum_{k=1}^G E(\beta_j | Y, y^{(k)}, \Omega^{(k)}, \tau^{(k)}) \right]^2$$

$$= \frac{1}{G} \sum_{k=1}^G (\tau_{(k)}^e X_T' X_T' X_T' + \Omega_{(k)}^{-1})^{-1} + \frac{1}{G} \sum_{k=1}^G (\tau_{(k)}^e X_T' X_T' X_T' + \Omega_{(k)}^{-1})^{-1} \times (\tau_{(k)}^e X_T' X_T' X_T' + \Omega_{(k)}^{-1})^{-1} \times \left[\frac{1}{G} \sum_{k=1}^G (\tau_{(k)}^e X_T' X_T' X_T' + \Omega_{(k)}^{-1})^{-1} (\tau_{(k)}^e X_T' X_T' X_T' + \Omega_{(k)}^{-1})^{-1} \right]$$

A2:

Lemme: $[\eta | X, \beta, \tau, \gamma, \Omega, \lambda]$ est une fonction concave logarithmique de η .

Preuve: Soit $h(\eta) = \log[\eta | X, \beta, \tau, \gamma, \Omega, \lambda]$. Il est suffisant de montrer que

$$\frac{\partial^2 h(\eta)}{\partial^2 \eta} \leq 0.$$

Manifestement,

$$\frac{\partial h(\eta)}{\partial \eta} = -m \frac{\Gamma'(\eta)}{\Gamma(\eta)} + m \log(\lambda) + \log(\prod_{i=1}^m \tau_i).$$

Soit $\psi(\eta) = \Gamma'(\eta)/\Gamma(\eta)$, nous avons alors

Le tableau 6 présente les estimations a posteriori de σ_j^2 à l'aide des différentes distributions a priori sur σ_j^2 . Comme l'indique le tableau 6, lorsque a_j et b_j sont petites ($\leq 0,01$), il n'existe pratiquement pas de différence entre les estimations. À mesure que a_j et b_j augmentent, les estimations σ_j^2 (RB) deviennent plus petites. Toutefois, en présence de solides informations a priori sur a_j et b_j , par exemple $a_j = b_j = 10$, les estimations a posteriori de σ_j^2 sont appréciablement différentes de celles qui sont obtenues pour des distributions a priori non informatives.

Tableau 6
Comparaison de variances de l'erreur
d'échantillonnage estimée

Région	IG(a_j, b_j), $a_j = b_j$					
	0,0001	0,001	0,01	0,1	1	10
1	40,09	40,10	40,05	39,64	37,14	22,29
2	34,19	34,18	34,17	33,97	31,74	19,05
3	94,48	94,49	94,42	93,76	86,73	50,60
4	52,08	52,08	52,04	51,63	48,21	28,82
5	121,60	121,70	121,60	121,40	113,70	66,75
6	94,03	94,03	93,83	92,96	87,21	52,90
7	102,30	102,30	102,20	101,40	94,85	57,58
8	160,10	160,00	159,90	159,10	147,60	86,61
9	63,46	63,46	63,38	62,99	58,46	34,85
10	65,88	65,87	65,89	65,40	60,76	36,60

4. CONCLUSION

Dans le présent exposé, nous avons présenté des méthodes bayésiennes hiérarchiques pour l'estimation régionale, à l'aide de modèles à plusieurs niveaux. Il n'est claire-ment pas facile de fournir un modèle approprié pour toutes les régions de façon à obtenir des résultats satisfaisants, même si des méthodes bayésiennes de type MCMC (Monte Carlo à chaîne markovienne) comme l'échantillonnage de Gibbs nous permettent d'ajuster les données à l'aide de modèles bayésiens d'une complexité virtuellement illimitée. La taille et l'homogénéité des régions et la présence de solides informations auxiliaires auront un effet sur le résultat final. Des modèles appropriés dans certaines situations pourront ne pas se prêter à d'autres situations. La méthode bayésienne hiérarchique comporte également des limitations, par exemple le choix de distributions a priori sur les paramètres de modèle et certaines questions d'échantillonnage liées à la méthode d'échantillonnage de Gibbs. Néanmoins, la méthode bayésienne hiérarchique générale s'applique à tout un choix de situations pour l'estimation de paramètres régionaux. Le choix du modèle est un aspect important de l'analyse bayésienne hiérarchique. Il est également important de comparer la méthode bayésienne hiérarchique à d'autres méthodes largement utilisées pour l'estimation régionale, par exemple la méthode bayésienne empirique et la méthode du meilleur prédicteur linéaire sans biais empirique. Les travaux se poursuivent afin d'y

faible que l'estimateur empirique $\mu_{(E)}^{(RB)}$ pour toutes les régions. Dans tous les cas, l'erreur-type de $\mu_{(E)}^{(RB)}$ représente de 50% à 75% environ de l'erreur-type de $\mu_{(E)}^{(E)}$, ce qui confirme l'avantage de la Rao-Blackwellisation. Ainsi, $\mu_{(RB)}^{(E)}$ est plus stable que $\mu_{(E)}^{(E)}$ lorsqu'on l'utilise pour obtenir des estimations ponctuelles pour les moyennes à posteriori dans une analyse bayésienne computationnelle. Il convient de mentionner que l'erreur-type de simulation de $\mu_{(E)}^{(E)}$ est également un estimateur de l'erreur-type à posteriori. Ainsi, l'erreur-type de simulation de $\mu_{(E)}^{(RB)}$ au tableau 3 est presque identique à l'erreur-type estimée $\mu_{(RB)}^{(E)}$ au tableau 2.

Tableau 3

Erreurs-types de simulation	
Région	$\mu_{(E)}^{(RB)}$
1	0,817
2	0,862
3	1,090
4	1,101
5	1,583
6	0,930
7	0,978
8	1,208
9	0,997
10	0,869
	0,513

Dans le modèle 2, les variances d'erreur $\tau_i = \sigma_i^2$ sont supposées indépendantes avec des distributions à priori $G(a_i, b_i)$ ou σ_i^2 avec la distribution gamma inverse $IG(a_i, b_i)$, où a_i et b_i sont des valeurs connues choisies de façon à refléter nos connaissances à priori de σ_i^2 . Dans la pratique, il est toujours difficile d'obtenir des informations exactes au sujet des variances d'échantillonnage. De plus, à mesure que le nombre de régions m augmente, le nombre de composantes des variances σ_i^2 augmente. Nous nous intéressons aux effets possibles du choix des distributions à priori sur les σ_i^2 ; en particulier, nous aimerions évaluer la sensibilité des moyennes à posteriori au choix des distributions à priori sur les σ_i^2 . Afin de vérifier la sensibilité des estimations à posteriori au choix de a_i et b_i dans le cadre du modèle 2, nous avons fixé six valeurs différentes de $a_i = b_i$, c'est-à-dire 0,0001, 0,001, 0,01, 0,1, 1 et 10.

$$\{\tau_i | Y, \beta, \gamma, \Omega\} \sim G\left(a_i + \frac{2}{n_i}b_i + \frac{1}{2}(Y_i - X_i'\beta)^\tau (Y_i - X_i'\beta)\right), \quad (19)$$

les effets d'échantillon $n_i^{1/2}$ et $(Y_i - X_i'\beta)^\tau (Y_i - X_i'\beta)/2$ dominent l'information à priori a_i et b_i lorsque a_i et b_i

Tableau 5

Comparaison de variances à posteriori de l'erreur d'échantillonnage estimée

Région		$IG(a_i, b_i), a_i = b_i$	
1	0,658	0,658	0,656
2	0,724	0,724	0,711
3	1,167	1,167	1,161
4	1,220	1,220	1,217
5	2,455	2,455	2,462
6	0,871	0,870	0,830
7	0,933	0,933	0,931
8	1,418	1,417	1,375
9	1,015	1,014	1,011
10	0,760	0,760	0,750
			0,745
			0,613

Le tableau 4 indique clairement que les estimations de moyennes régionales sont très stables; elles ne sont pas sensibles au choix de a_i et b_i . Toutefois, comme le montre le tableau 5, les variances à posteriori diminuent à mesure que les distributions à priori sur σ_i^2 deviennent plus informatives, et même à de plus petits coefficients de variation (cv). Cela indique que nous pouvons améliorer les résultats des estimations pour les régions en ce qui concerne le cv si nous avons plus de renseignements à priori sur les variances de l'erreur d'échantillonnage. Dans notre étude, nous avons considéré uniquement le cas $a_i = b_i$. Une étude plus approfondie comporterait différentes combinaisons de a_i et b_i .

Tableau 4

Comparaison de moyennes régionales estimées

Région		$IG(a_i, b_i), a_i = b_i$	
1	10,23	10,23	10,25
2	9,84	9,84	9,82
3	13,00	13,00	13,07
4	10,95	10,95	10,94
5	17,86	17,87	17,78
6	10,21	10,21	10,25
7	9,58	9,58	9,63
8	10,29	10,30	10,37
9	11,34	11,34	11,32
10	9,79	9,79	9,82
			9,92

sont petites. Ainsi $IG(0,0001, 0,0001)$, $IG(0,001, 0,001)$ et $IG(0,01, 0,01)$ peuvent être considérées comme des distributions à priori non informatives tandis que $IG(1, 1)$ et $IG(10, 10)$ peuvent être considérées comme des distributions à priori informatives. Le tableau 4 présente des moyennes à posteriori dans le cadre du modèle 2 à l'aide des différentes distributions à priori sur σ_i^2 , et le tableau 5 montre les variances à posteriori correspondantes.

avons examinée la valeur de $f(y_{ij} | x_{(ij)})$ pour la donnée simple observée, la soi-disant ordonnée prédictive conditionnelle (CPO) pour chacun des trois modèles. Nous avons

$$CPO_{ij} = f(y_{ij}^{obs} | x_{(ij)}^{obs})$$

où y_{ij}^{obs} désigne la donnée simple observée. Puisque les CPO ne sont que les vraisemblances observées, les modèles comportant des CPO plus élevées fournissent un meilleur ajustement pour les données observées. À l'aide de la sortie de l'échantillonneur de Gibbs, nous pouvons calculer les CPO pour toutes les données simples. Ainsi, dans le cadre du modèle 1, nous avons

$$\frac{f(y_{ij} | x_{(ij)})}{f(x_{(ij)})}$$

$$= \frac{\int f(x_{(ij)} | y_{ij}, \beta, \sigma^2) \cdot f(\beta, \sigma^2 | x) d\beta d\sigma^2}{1} = \frac{\int \frac{f(y_{ij} | x_{(ij)}, \beta, \sigma^2)}{1} \cdot f(\beta, \sigma^2 | x) d\beta d\sigma^2}{1}$$

Nous notons maintenant que les y_{ij} sont conditionnellement indépendantes, donc $f(y_{ij} | x_{(ij)}, \beta, \sigma^2) = f(y_{ij} | \beta, \sigma^2)$, et les valeurs CPO sont calculées comme suit:

$$\widehat{CPO}_{ij} = \frac{1}{1} \cdot \frac{1}{\sum_{k=1}^G G(y_{ij}^{obs} | \beta_{(k)}', \sigma_{2(k)}^2)}$$

où $f(y_{ij} | \beta, \sigma^2)$ est la fonction de densité normale donnée par (3). Pour les modèles 2 et 3, on calcule les CPO avec $\sigma_{2(k)}^2$ remplacée par $\sigma_{1(k)}^2$ en (18). On trouvera une discussion plus détaillée dans Gelfand (1995). Nous présentons un tracé CPO pour les trois modèles à la figure 1. Le modèle 2 est clairement le meilleur, puisqu'il y a une majorité de valeurs CPO pour le modèle 2 est appréciablement plus grande que celles des modèles 1 et 3. Le modèle 3 est légèrement meilleur que le modèle 1 pour ce qui est des valeurs CPO. Il y a aussi de petites valeurs CPO pour les trois modèles, ce qui indique que les hypothèses de nos modèles ne sont peut-être pas bien respectées par notre ensemble de données.

Compte tenu des estimations de la variance d'échantillonnage que l'on trouve dans le tableau 1 et le tracé CPO, nous concluons que le modèle 2 est le meilleur modèle d'ajustement pour nos données. Par conséquent, nous avons utilisé le modèle 2 pour obtenir des estimations modélisées de moyennes régionales et des erreurs-types a posteriori connexes.

3.3 Résultat de l'estimation

Nous présentons maintenant les estimations des moyennes régionales en fonction du modèle 2 seulement. Le tableau 2 montre les moyennes régionales a posteriori estimées et les erreurs-types a posteriori correspondantes. Notre étude a montré que l'estimateur empirique $\hat{\mu}_{(E)}^{(RB)}$ de l'estimateur Rao-Blackwellisé $\hat{\mu}_{(RB)}^{(E)}$ fournissent presque les mêmes estimations ponctuelles, et nous avons donc signalé uniquement les estimations obtenues à l'aide de $\hat{\mu}_{(RB)}^{(E)}$. Pour la comparaison, nous avons calculé les estimations directes (moyennes d'échantillon) et les erreurs-types directes correspondantes pour les 10 régions. Le tableau 2 montre clairement que les estimations modélisées sont appréciablement plus efficaces que les estimations directes. Les erreurs-types a posteriori sont beaucoup plus petites que les erreurs-types directes.

Tableau 2

Estimations de moyennes régionales		
Région	$\hat{\mu}_{(E)}^{(RB)}$	$\hat{\mu}_{(E)}^{(RB)}$
1	11,08	9,53
2	7,91	6,82
3	13,48	14,15
4	6,53	8,01
5	19,52	14,96
6	11,21	11,38
7	8,72	11,24
8	12,81	13,99
9	10,18	8,76
10	10,01	11,30
	9,79	11,34
	1,01	1,19
	0,97	0,93
	1,57	1,11
	1,08	1,05
	0,85	0,81

Afin d'étudier les effets de la Rao-Blackwellisation, nous avons calculé les erreurs-types de simulation de $\hat{\mu}_{(E)}^{(RB)}$, qui sont respectivement les erreurs-types d'échantillon de $\{X_{1T}'\beta_{(k)}', y_{(k)}', \Omega_{(k)}', \tau_{(k)}'\}$. Le tableau 3 montre les erreurs-types de simulation. Le tableau 3 indique clairement que l'estimateur Rao-Blackwellisé $\hat{\mu}_{(RB)}^{(E)}$ comporte une erreur-type de simulation beaucoup plus

A la section 2, nous avons proposé trois modèles en fonction d'hypothèses différentes quant aux variances d'échantillonnage. Afin de déterminer quel modèle ajustait le mieux nos données, nous avons d'abord obtenu les estimations a posteriori des variances d'échantillonnage dans le cadre des trois modèles. Nous avons également calculé les estimations des moindres carrés ordinaires (MCO) des variances d'échantillonnage au sein de chaque région à l'aide des données propres à la région seulement. Le tableau 1 indique les estimations Rao-Blackwellisées des variances d'échantillonnage dans le cadre des trois modèles de même que les estimations MCO.

Variances estimatives de l'erreur d'échantillonnage

Dans le tableau 1, les estimations MCO indiquent de fortes variations parmi les 10 régions. Le modèle 1 suppose

Afin d'examiner comment les données appuient chaque modèle, nous avons calculé la densité prédictive de validation croisée pour chaque donnée simple y^j . La densité de validation croisée pour y^j est la densité conditionnelle $f(y^j | X^{(j)})$, où $X^{(j)}$ désigne toutes les données sauf y^j . Nous

2 pour notre analyse des données. De vagues distributions a priori appropriées sur des paramètres inconnus sont supposées être des variables normales indépendantes de moyenne 0 et d'écart-type 100, de sorte qu'un intervalle a priori de 95% correspond à $\pm 2\sigma$ environ, et la distribution a priori est uniforme localement pour la région appuyée par la vraisemblance. Comme autre possibilité, une distribution a priori uniforme sur un intervalle convenablement large pourrait être donnée, par exemple $U(-200, 200)$. Une distribution a priori de Wishart $W_3^-(\alpha, R)$ est spécifiée pour la matrice de covariances inverse $\Omega = \Phi^{-1}$. Pour représenter une vague connaissance a priori, nous avons choisi les degrés de liberté α pour que cette distribution soit aussi petite que possible, $\alpha = 3$, le rang de Ω (Spiegelhalter, Thomas, Best et Gilks 1996). La matrice scalaire R est décrite avec des éléments diagonaux égaux à 1 et des éléments non diagonaux égaux à 0,001, ce qui représente notre estimation préliminaire pour l'ordre de grandeur de la matrice de covariances. Pour les modèles 1 et 2, une distribution a priori gamma $G(0,001, 0,001)$ est supposée pour τ_0 et les τ_i . Pour le modèle 3, $\tau_1 \sim G(\eta, \lambda)$, et on suppose que η et λ sont distribués indépendamment comme $U(0, 10000)$, c'est-à-dire la distribution uniforme sur un grand intervalle. Nous nous attendons à ce que les vagues distributions a priori appropriées sur les hyperparamètres se rapprochent raisonnablement bien des distributions a priori uniformes et qu'elles exercent ainsi un effet

Nous cherchons à estimer la moyenne régionale $\mu_1 = \frac{X_1'}{X_1} \bar{b}_1 = \beta_0 + \frac{X_1'}{X_1} \bar{b}_{11} + \frac{X_2'}{X_1} \bar{b}_{12}$, où $\frac{X_1'}{X_1}$ et $\frac{X_2'}{X_1}$ sont les moyennes de population de la i -ième région des variables auxiliaires x_1 et x_2 , respectivement. Pour ce faire, nous allons d'abord choisir un modèle optimal pour l'ensemble de données, et nous allons présenter les estimations modélisées pour les moyennes régionales fondées sur le modèle choisi.

(iv) Distributions a priori marginales : $\gamma \sim N^p(0, D)$, $\Omega \sim W^p(\alpha, R)$, $\eta \sim U^+$ et $\lambda \sim U^+$, où U^+ désigne une distribution uniforme sur un sous-ensemble de R^+ avec une longueur grande mais finie, D , α et R connus.

Remarque 3.1 : Dans le modèle 3, nous supposons que les r_i sont des variables aléatoires gamma indépendantes et distribuées de façon identique avec des hyperparamètres η et λ inconnus. Nous avons ainsi des modèles de population pour la variance d'échantillonnage σ_i^2 , aussi bien que pour la variance d'échantillonnage σ_i^2 . Dans les modèles 1 et 2, nous avons considéré la modélisation de β_i seulement et nous avons supposé de vagues distributions a priori appropriées sur σ_i^2 ou σ_i^2 .

Remarque 3.2 : L'hypothèse (iii) n'est peut-être pas un bon modèle de population pour tous les r_i . Comme autre possibilité, nous pouvons modéliser r_i de façon plus réaliste, comme dans le cas de β_i en précisant un modèle de régression pour le logarithme de r_i . Cela peut exiger des renseignements auxiliaires liés à r_i . Dans l'analyse des données de la section 3, toutefois, nous utilisons simplement $G(\eta, \lambda)$ comme modèle de population pour r_i . Il n'est généralement pas facile de modéliser les variances d'échantillonnage lorsqu'elles sont inconnues.

Les distributions conditionnelles complètes pour un échantillonnage de Gibbs dans le cadre du modèle 3 sont données par :

$$(i) \quad \text{Pour } i = 1, \dots, m, \quad [\beta_i | r_i, \tau_i, \gamma, \Omega, \eta, \lambda] \sim \text{ind } N^p((r_i' X_i' X_i' + \Omega)^{-1} (r_i' X_i' \tau_i + \Omega Z_i' \gamma), (r_i' X_i' X_i' + \Omega)^{-1})$$

$$(ii) \quad \text{Pour } i = 1, \dots, m, \quad [r_i | r_i, \beta_i, \gamma, \Omega, \eta, \lambda] \sim \text{ind } G\left(\eta + \frac{2}{n_i}, \frac{1}{2}(X_i' - X_i' \beta_i)'(X_i' - X_i' \beta_i) + \lambda\right)$$

$$(iii) \quad [\gamma | r_i, \beta_i, \tau_i, \eta, \lambda] \sim N^p\left(\left(\sum_{i=1}^m r_i' \Omega Z_i' + D\right)^{-1} \left(\sum_{i=1}^m r_i' \Omega \beta_i' + Z_i' \tau_i\right), \left(\sum_{i=1}^m r_i' \Omega Z_i' + D\right)^{-1}\right)$$

$$(iv) \quad [\Omega | r_i, \beta_i, \sigma_i^2, \gamma, \eta, \lambda] \sim W^p$$

$$(v) \quad [\eta | r_i, \beta_i, \tau_i, \gamma, \Omega, \lambda] \propto [T(\eta)]^{-m} \lambda_m (I_m^{-1} \tau_i)$$

$$(vi) \quad [\lambda | r_i, \beta_i, \tau_i, \gamma, \Omega, \eta] \sim G(m\eta + 1, \sum_{i=1}^m \tau_i)$$

Pour le modèle 3, les estimateurs a posteriori de β_i , τ_i et λ_i ont les mêmes formes que celles qui ont été données pour le modèle 2. Dans le cadre du modèle 3, l'estimateur Rao-Blackwellisé de la moyenne a posteriori de σ_i^2 est donné par

Dans le cadre du modèle 3, $[\eta | r_i, \beta_i, \tau_i, \gamma, \Omega, \lambda]$ est connu jusqu'à une constante multiplicatrice seulement. Toutefois, puisque $[\eta | r_i, \beta_i, \tau_i, \gamma, \Omega, \lambda]$ est une fonction concave logarithmique de η (voir l'annexe A2), il est possible d'utiliser la méthode d'échantillonnage à rejet adaptée de Gilks, Best et Tan (1995) dans l'échantillonneur de Gibbs afin de produire des échantillons à partir de la distribution conditionnelle $[\eta | r_i, \beta_i, \tau_i, \gamma, \Omega, \lambda]$.

3. ANALYSE DES DONNÉES

3.1 Description des données et des modèles

Suivant Holt et Moura (1993) et Moura et Holt (1999), nous avons considéré l'estimation du revenu des ménages de comtés (régions) du Brésil. Les données originales de Holt et Moura comportent 140 régions, les unités d'échantillonnage étant tirées de chaque région par échantillonnage aléatoire simple. Par conséquent, nous avons utilisé uniquement une faible partie de l'ensemble original de données sans notre analyse, à simple titre d'illustration. Notre ensemble de données contient un sous-ensemble de 10 régions comportant 28 unités d'échantillonnage obtenues par échantillonnage aléatoire simple dans chaque région.

Soit y_{ij} , le revenu du j -ième ménage dans la i -ième région. Il y a deux variables auxiliaires au niveau de l'unité, c'est-à-dire x_1 et x_2 , où x_1 désigne le nombre de pièces dans un ménage et x_2 désigne le niveau de scolarité atteint par le chef du ménage. Le modèle d'échantillonnage est

$$(17) \quad y_{ij} = x_{1ij} \beta_i + e_{ij} = \beta_{0i} + x_{1ij} \beta_{1i} + x_{2ij} \beta_{2i} + e_{ij},$$

où x_{1ij} désigne le nombre de pièces dans le j -ième ménage de la région i et x_{2ij} désigne le niveau de scolarité correspondant à l'erreur global respectives et e_{ij} est la variable de l'échantillon global respectives et e_{ij} est la variable de l'erreur d'échantillonnage, sa distribution étant précisée par les trois modèles de la variance d'erreur discutées à la section 2. Dans le modèle d'échantillonnage (17), β_i est le coefficient de régression aléatoire correspondant à la i -ième région et se laisse modéliser comme

$\beta_{0i} = \gamma_0 + v_{0i}$, $\beta_{1i} = \gamma_{10} + v_{1i}$, $\beta_{2i} = \gamma_{20} + v_{2i}$, où $\gamma = (\gamma_0, \gamma_{10}, \gamma_{20})^T$ est le vecteur inconnu de paramètres de régression fixes, $v_i = (v_{0i}, v_{1i}, v_{2i})^T$ est le vecteur d'effets aléatoires de la i -ième région distribué comme

$$V(\mu_i) = \overline{X}_i' V(\beta) \overline{X}_i \quad (10)$$

On peut appliquer la même procédure d'estimation à la variance de l'erreur d'échantillonnage σ_2^2 . Puisque, conditionnellement, σ_2^2 comporte une distribution gamma inverse, l'estimateur Rao-Blackwellisé de la moyenne a posteriori de σ_2^2 s'obtient sous la forme

$$\hat{\sigma}_{2(RB)}^2 = \frac{1}{G} \sum_{i=1}^G \left(b + \frac{1}{m} \sum_{j=1}^m (Y'_i - X'_j \beta_{(j)}^T)(Y'_i - X'_j \beta_{(j)}^T) \right) \quad (11)$$

Puisque nous intéressons surtout à l'estimation des moyennes régionales, le calcul de la variance d'échantillonnage sert uniquement à la sélection du modèle. On trouvera des détails sur la sélection du modèle à la section 3.2.

2.2 Variance de l'erreur inégale

Dans la pratique, il est plus réaliste d'admettre une variance d'erreur inégale pour les erreurs d'échantillonnage. Soit σ_j^2 , la vraie variance de l'erreur d'échantillonnage pour la i -ième région. Une extension directe du modèle 1 mène au modèle bayésien hiérarchique à plusieurs niveaux de la variance d'erreur inégale que voici:

Modèle 2:

- (i) Sous réserve que β_i et σ_j^2 , les y_{ij} sont indépendants avec
- $$y_{ij} | \beta_i, \sigma_j^2 \sim N(x_{ij}^T \beta_i, \sigma_j^2), \quad (i = 1, \dots, m; j = 1, \dots, n_i); \quad (12)$$
- (ii) Sous réserve que γ et Φ , les β_i sont indépendants avec
- $$\beta_i | \gamma, \Phi \sim N^p(Z_i' \gamma, \Phi), \quad (i = 1, \dots, m); \quad (13)$$
- (iii) Distributions a priori marginales: $\gamma \sim N^q(0, D)$, $\Omega \sim W_p^q(a, R)$, où $\tau_i \sim G(a_i, b_i)$ et D, a_i, b_i, a et R connus.

Remarque 2.1: Le modèle 2 se réduit au modèle 1 lorsque $\sigma_2^2 = \sigma_2^2$ pour tous les i . D'un point de vue hiérarchique de Bayes, l'extension du modèle de la variance d'erreur égale au modèle de la variance d'erreur inégale va de soi. Il n'y a pas non plus de difficulté pour la mise en oeuvre de l'échantillonnage de Gibbs.

Remarque 2.2: Les τ_i sont supposés indépendants et comportent des distributions a priori $G(a_i, b_i)$, où a_i et b_i sont des hyperparamètres connus normalement choisis très petits afin de refléter une vague connaissance des τ_i . Les distributions conditionnelles complètes pour un échantillonnage de Gibbs dans le cadre du modèle 2 sont données par:

$$\hat{\sigma}_{2(RB)}^2 = \frac{1}{G} \sum_{i=1}^G \left[b_i + \frac{1}{n_i} (Y'_i - X'_j \beta_{(j)}^T)(Y'_i - X'_j \beta_{(j)}^T) \right] \times \left(a_i + \frac{1}{n_i} n_i - 1 \right)^{-1}. \quad (14)$$

2.3 Variance de l'erreur aléatoire

Dans le modèle 2, nous avons supposé une variance d'erreur inégale pour les erreurs d'échantillonnage. Kiefer et Rao (1992) ont utilisé un modèle simple de la variance d'erreur aléatoire afin d'obtenir les meilleurs prédicteurs linéaires sans biais pour des moyennes régionales. Dans la présente section, nous étendons leur modèle au cas comportant plusieurs niveaux. Nous supposons des modèles d'effets aléatoires sur les coefficients de régression β_i aussi bien que sur les variances de l'erreur d'échantillonnage σ_j^2 , ce qui mène au modèle 3 donné ci-dessous.

Modèle 3:

- (i) Comme pour le modèle 2;
- (ii) Comme pour le modèle 2;
- (iii) Sous réserve de η et de λ , les τ_i sont indépendants avec
- $$\tau_i | \eta, \lambda \sim G(\eta, \lambda), \quad \text{où } \tau_i = \sigma_j^{-2}; \quad (15)$$

Nous voulons trouver les distributions a posteriori des β_j pour les données $X = \{\gamma_j, i = 1, \dots, m; j = 1, \dots, n\}$ et, en particulier, trouver les estimations a posteriori des moyennes régionales $\mu_j = X_j/\beta_j$, qui dépendent des estimations de β_j . L'évaluation directe de la distribution a posteriori composée comporte une intégration numérique très dimensionnelle, et le calcul n'est pas faisable. Par conséquent, nous utilisons la méthode d'échantillonnage de Gibbs (Gelfand et Smith 1990) afin de produire des échantillons à partir des distributions a posteriori composées. Pour appliquer l'échantillonnage de Gibbs dans le cadre du modèle 1, il nous faut les distributions conditionnelles complètes données par :

(i) Pour $i = 1, \dots, m$,

$$[\beta_j | X, \gamma, \Omega, \tau_j] \propto N^d(\tau_j^e X_T' X_T' + \Omega^{-1} \\ (\tau_j^e X_T' Y_j + \Omega Z_T' X_T' + \Omega) - 1)$$

$$[\gamma | X, \beta, \Omega, \tau_j] \propto N^b \left(\left(\sum_{i=1}^n Z_i' Z_i' \Omega Z_i' + D^{-1} \right)^{-1} \left(\sum_{i=1}^n Z_i' \Omega \beta_j + \left(\sum_{i=1}^n Z_i' Z_i' \Omega Z_i' + D^{-1} \right)^{-1} \right) \right)$$

$$[\Omega | X, \beta, \gamma, \tau_j] \propto W^d \left(\alpha + m, R + \frac{1}{2} \sum_{i=1}^n (\beta_j' - Z_i' \gamma) (\beta_j' - Z_i' \gamma)' \right)$$

$$[\tau_j^e | X, \beta, \gamma, \Omega] \propto G \left(a + \frac{1}{2} \sum_{i=1}^n n_i, b + \frac{1}{2} \sum_{i=1}^n (\gamma_i' - X_i' \beta_j) (\gamma_i' - X_i' \beta_j)' \right)$$

Toutes les distributions conditionnelles complètes sont échantillons. La méthode d'échantillonnage de Gibbs se présente comme suit : a) À l'aide de valeurs de départ $\gamma^{(0)}$, $\Omega^{(0)}$ et $\tau_j^{(0)}$, il s'agit de tirer $\beta_j^{(1)}$, $i = 1, \dots, m$, à partir de $[\beta_j | X, \gamma, \Omega, \tau_j]$; b) de tirer $\gamma^{(1)}$ à partir de $[\gamma | X, \beta, \Omega, \tau_j]$ à l'aide de $\beta_j^{(1)}$, $i = 1, \dots, m$, $\Omega^{(0)}$ et $\tau_j^{(0)}$; c) de tirer $\Omega^{(1)}$ à partir de $[\Omega | X, \beta, \gamma, \tau_j]$ à l'aide de $\beta_j^{(1)}$ et $\tau_j^{(1)}$; d) de tirer $\tau_j^{(1)}$ à partir de $[\tau_j | X, \beta, \gamma, \Omega]$ à l'aide de $\beta_j^{(1)}$, $i = 1, \dots, m$, $\gamma^{(1)}$ et $\Omega^{(1)}$. Les étapes (a)-(d) permettent de compléter un cycle d'échantillonnage. On exécute un nombre élevé de cycles, disons t , période dite de rodage, jusqu'à la convergence, puis on traite $\{\beta_j^{(t+k)}, i = 1, \dots, m; \gamma^{(t+k)}, \Omega^{(t+k)}, \tau_j^{(t+k)}; k = 1, \dots, G\}$ comme des échantillons G à partir de la distribution a posteriori composée de β_j , $i = 1, \dots, m$, γ , Ω et τ_j . Supposons qu'un échantillon de taille G , de la forme $\{\beta_j^{(k)}, i = 1, \dots, m; \gamma^{(k)}, \Omega^{(k)}, \tau_j^{(k)}; k = 1, \dots, G\}$, on peut utiliser la moyenne d'échantillon du $\{\beta_j^{(k)}\}$. Puisque β_j comporte une distribution conditionnelle complète de forme fermée, nous pouvons utiliser la moyenne d'échantillon des espérances conditionnelles

$$\beta_j^{(k)} = \frac{1}{G} \sum_{k=1}^G \beta_j^{(k)} \quad (5)$$

et avons ainsi les deux autres estimateurs ci-dessous pour β_j : l'estimateur correspondant est l'estimateur soit-disant Rao-Blackwellisé (Gelfand et Smith 1990, 1991). Nous fonde sur le théorème bien connu de Rao-Blackwell, et l'estimation, puisque $E(\beta_j | X) = E(E(\beta_j | X, \gamma, \Omega, \tau_j))$, et $\text{Var}(\beta_j | X) \geq \text{Var}(E(\beta_j | X, \gamma, \Omega, \tau_j))$. Cette modification note $\{E(\beta_j | X, \gamma^{(k)}, \Omega^{(k)}, \tau_j^{(k)})\}$ afin d'améliorer notre

$$\beta_j^{(RB)} = \frac{1}{G} \sum_{k=1}^G E(\beta_j | X, \gamma^{(k)}, \Omega^{(k)}, \tau_j^{(k)})$$

$$\mu_j^{(RB)} = \frac{1}{G} \sum_{k=1}^G \tau_j^{(k)} X_T' Y_j + \Omega^{(k)} Z_T' X_T' Y_j \quad (6)$$

où $\beta_j^{(RB)}$ est l'estimateur empirique et $\beta_j^{(RB)}$ est l'estimateur Rao-Blackwellisé. $\beta_j^{(E)}$ et $\beta_j^{(RB)}$ sont tous deux sans biais pour la moyenne a posteriori. Toutefois, $\beta_j^{(RB)}$ est meilleur que $\beta_j^{(E)}$ pour ce qui est de l'erreur-type de simulation (Gelfand et Smith 1991).

Les estimateurs correspondants pour la moyenne régionale μ_j sont donnés comme suit

$$\mu_j^{(E)} = X_T' \beta_j^{(E)} = \frac{1}{G} \sum_{k=1}^G X_T' \beta_j^{(k)} \quad (7)$$

$$\mu_j^{(RB)} = \frac{1}{G} \sum_{k=1}^G X_T' \beta_j^{(RB)} = \frac{1}{G} \sum_{k=1}^G X_T' \tau_j^{(k)} X_T' Y_j + \Omega^{(k)} Z_T' X_T' Y_j \quad (8)$$

Nous nous attendons à ce que $\mu_j^{(E)}$ et $\mu_j^{(RB)}$ donnent tous deux presque les mêmes estimations ponctuelles. Toutefois, il est intéressant de calculer et de comparer les erreurs-types de simulation des effets de la Rao-Blackwellisation (voir la section 3).

Afin d'obtenir la variance a posteriori de μ_j , nous trouvons d'abord la variance a posteriori de β_j , puisque

$$V(\mu_j | X) = X_T' V(\beta_j | X) X_T'$$

$$V(\beta_j | X) = E(V(\beta_j | X, \gamma, \Omega, \tau_j)) + V(E(\beta_j | X, \gamma, \Omega, \tau_j))$$

$$- [E(E(\beta_j | X, \gamma, \Omega, \tau_j))]^2 \quad (9)$$

À l'aide de (9), l'estimateur Rao-Blackwellisé de la variance a posteriori de β_j , noté $V(\beta_j)$, s'obtient à l'aide des échantillons de Gibbs $\{\beta_j^{(k)}, i = 1, \dots, m; \gamma^{(k)}, \Omega^{(k)}, \tau_j^{(k)}; k = 1, \dots, G\}$; voir l'annexe A1. La variance a posteriori de la moyenne régionale μ_j est alors estimée à l'aide de

comités (régions) du Brésil, ils ont obtenu un accroissement de l'efficacité des estimateurs de type MPLSB empirique relatif aux estimateurs MPLSB empiriques obtenus à l'aide de modèles de régression à erreur emboîtée. Ghosh et Rao (1994) et Rao (1999) ont donné un aperçu détaillé des méthodes d'estimation régionale modélisées.

Dans le présent exposé, nous examinons le modèle (2) à plusieurs niveaux selon un cadre hiérarchique de Bayes et nous l'étendons à des modèles à plusieurs niveaux plus généraux qui admettent des variances de l'erreur inégale ou des variances de l'erreur aléatoires fixes. La moyenne régionale μ_j est estimée à l'aide de sa moyenne à posteriori, et sa précision est mesurée en fonction de sa variance à posteriori. La variance à posteriori tient compte automatiquement de l'incertitude supplémentaire associée aux hyperparamètres bayésiennes hiérarchiques et les variances à posteriori connexes. À la section 2, nous présentons les modèles bayésiens hiérarchiques à plusieurs niveaux avec diverses hypothèses pour ce qui est de la variance de l'erreur et de l'inférence de l'échantillonnage de Gibbs connexe. À la section 3, nous illustrons notre méthode et nous abordons la sélection du modèle et l'analyse de la sensibilité à l'aide de données sur le revenu des ménages de comtés (régions) du Brésil. Enfin, à la section 4, nous présentons des remarques et nos conclusions.

2. MODÈLE À PLUSIEURS NIVEAUX ET INFÉRENCE DE L'ÉCHANTILLONNAGE DE GIBBS

2.1 Variance de l'erreur égale

Nous considérons une représentation bayésienne hiérarchique du modèle (2) à plusieurs niveaux comme suit:

Modèle 1:

(i) Sous réserve que β_j et σ_j^2 , les y_{ij} sont indépendants avec

$$y_{ij} | \beta_j, \sigma_j^2 \sim N(x_{ij}^T \beta_j, \sigma_j^2),$$

(ii) $(i = 1, \dots, m; j = 1, \dots, n_j)$;

(iii) Sous réserve que γ et Φ , les β_j sont indépendants avec

$$\beta_j | \gamma, \Phi \sim N^p(Z_j^T \gamma, \Phi), (j = 1, \dots, m). \quad (4)$$

À fin de compléter notre description de modèle bayésien, nous adaptons les distributions à priori pour des paramètres comme suit:

(iii) Distributions à priori marginales: $\gamma \sim N^p(0, D)$, $\tau_g \sim G(a, b)$ et $\Omega \sim W^p(\alpha, R)$, où $\tau_g = \sigma_g^2$, $\Omega = \Phi^{-1}$ et D, a, b, α et R connus.

À l'étape (iii) du modèle 1, $G(a, b)$ désigne une distribution gamma avec une densité donnée par

$$f(X) = \frac{|R|^{\frac{a}{2}}}{2^{\frac{a}{2}} \Gamma(\frac{a}{2})} |X|^{\frac{a}{2}} \exp\left\{-\frac{1}{2} \text{tr}(RX)\right\},$$

où $X > 0$, $R > 0$ et $\Gamma_p(\alpha)$ est une fonction gamma à plusieurs variables définie comme

$$\Gamma_p(\alpha) = \pi^{\frac{p(p-1)}{4}} \prod_{j=1}^p \Gamma\left(\alpha + \frac{1}{2}(1-j)\right).$$

Remarque 1.1: Les distributions à priori à l'étape (iii) et les distributions d'échantillonnage et de population données par (3) et par (4) sont conjuguées en ce sens qu'elles mènent à des distributions conditionnelles complètes pour γ , τ_g et Ω qui sont encore une fois des distributions normales conditionnelles complètes pour un paramètre quelconque sera connue jusqu'aux constantes normalisatrices; dans ce cas, une génération aléatoire plus perfectionnée sera requise; (2) des distributions conditionnelles complètes de forme fermée peuvent servir à trouver les estimateurs Rao-Blackwellisés des moyennes à posteriori et des variances à posteriori, et donc à améliorer l'estimation à posteriori. En général, pour une inférence bayésienne, le choix de distributions à priori n'est pas une tâche simple puisque toute distribution à priori appropriée sur les paramètres de modèle est un candidat plausible. C'est là une limitation des méthodes bayésiennes.

Remarque 1.2: Il importe de noter que nous avons utilisé des distributions à priori appropriées sur tous les paramètres inconnus afin de nous assurer que toutes les distributions à posteriori sont appropriées (Hobert et Casella 1996). Ainsi, nous n'avons pas à craindre la présence de distributions à posteriori inappropriées. La valeur des paramètres des distributions à priori (les hyperparamètres) est choisie de façon à refléter une connaissance assez vague des distributions à priori. Des détails seront fournis à la section 3 sur l'analyse des données.

Remarque 1.3: Dans le modèle 1, nous supposons une variance de l'erreur égale σ_g^2 pour toutes les régions. Dans la pratique, toutefois, les variances de l'erreur d'échantillonnage pourraient être différentes pour diverses régions. Un modèle plus général devrait admettre des variances d'erreur possiblement différentes. Dans les sections 2.2 et 2.3, nous introduirons des modèles de variance d'erreur inégale et de variance d'erreur aléatoire.

Estimation bayésienne hiérarchique des moyennes pour petites régions à l'aide de modèles à plusieurs niveaux

YONG YU et J.N.K. RAO¹

RÉSUMÉ

On examine les modèles standard à plusieurs niveaux comportant des paramètres de régression aléatoires pour les estimations régionales. Les auteurs élargissent également les modèles en admettant une variance intégrale de l'erreur ou en supposant des modèles à effets aléatoires tant pour les paramètres de régression que pour la variance de l'erreur. Les auteurs présentent ces modèles en fonction d'un cadre hiérarchique de Bayes et ils estiment une moyenne régionale en fonction de sa moyenne a posteriori. La variance a posteriori de la moyenne régionale sert de mesure de la précision de l'estimation. Elle tient compte automatiquement de l'incertitude supplémentaire associée aux hyperparamètres du modèle à plusieurs niveaux. Les auteurs utilisent l'échantillonnage de Gibbs pour calculer les moyennes a posteriori et la variance a posteriori des moyennes régionales. Ils obtiennent des estimateurs Rao-Blackwellisés réduisant les erreurs de Monte Carlo. On étudie également la sélection d'un modèle bayésien et l'analyse de la sensibilité. La procédure est illustrée à l'aide de données sur le revenu des ménages de quelques comtés (régions) du Brésil.

MOTS CLÉS : Échantillonnage de Gibbs; hiérarchique de Bayes; modèle à plusieurs niveaux; variance de l'erreur d'échantillonnage; région.

1. INTRODUCTION

Depuis quelques années, on accorde beaucoup d'attention à l'estimation régionale à cause de la demande accrue d'estimateurs régionaux fiables. Les estimateurs directs traditionnels propres à une région ne fournissent pas une précision adéquate puisque la taille des échantillons régionaux est rarement assez grande. Il est donc nécessaire d'utiliser des estimateurs indirects qui s'appuient sur des régions connexes, notamment des estimateurs indirects modèles. Battese, Harter et Fuller (1988) ont proposé et appliqué un modèle de régression à erreur emboîtée afin d'obtenir des estimations régionales modélisées. Le modèle se présente

$$y_{ij} = x_{ij}^T \beta + v_{0i} + e_{ij}, j = 1, \dots, n_i; i = 1, \dots, m, \quad (1)$$

comme suit:

où y_{ij} représente les observations associées aux unités échantillonnées de la i -ième région, $i = 1, \dots, m$, représente le vecteur $p \times 1$ des variables explicatives au niveau de l'unité, β est un ensemble de p paramètres de régression fixes, v_{0i} sont des effets régionaux indépendants avec $E(v_{0i}) = 0$ et $V(v_{0i}) = \sigma_v^2$. Nous supposons que les e_{ij} sont des variables de l'erreur aléatoire indépendantes avec $E(e_{ij}) = 0$ et $V(e_{ij}) = \sigma_e^2$. v_{0i} et e_{ij} sont également supposés indépendants. Pour l'ensemble de la population, le modèle (1) s'applique avec n_i remplacé par N_i , la taille de la population régionale. Le modèle (1) se laisse exprimer en fonction d'une notation matricielle comme suit:

$$Y_i = X_i' \beta + v_{0i} + e_i, i = 1, \dots, m,$$

où $Y_i = (y_{i1}, \dots, y_{in_i})^T$, $X_i = (x_{i1}, \dots, x_{in_i})^T$ est une matrice $n_i \times p$, $1_{n_i} = (1, \dots, 1)^T$ est le vecteur unité de longueur n_i et $e_i = (e_{i1}, \dots, e_{in_i})^T$.
Holt et Moura (1993) ont élargi le cadre ci-dessus pour en faire un modèle à plusieurs niveaux en introduisant des coefficients de régression aléatoires et en les rattachant à des variables explicatives au niveau de la région afin d'expliquer une partie des variations entre les régions. Le modèle se laisse énoncer comme suit:

$$Y_i = X_i' \beta_i + e_i, i = 1, \dots, m, \quad (2)$$

où Z_i est la matrice du plan $p \times q$ des variables au niveau de la région, γ est un vecteur $q \times 1$ de coefficients fixes et $v_i = (v_{i1}, \dots, v_{id_i})^T$ est un vecteur $p \times 1$ d'effets aléatoires pour la i -ième région. Les v_i sont indépendants d'une région à l'autre et comportent une distribution composée au sein de chaque région avec $E(v_i) = 0$ et $V(v_i) = \Phi$, où la matrice de variances-covariances Φ est inconnue. À noter que le modèle (2) intègre effectivement le recours à des covariables au niveau de l'unité et au niveau de la région en un même modèle. Holt et Moura (1993) et Moura et Holt (1999) ont élargi le cadre adopté par Prasad et Rao (1990) pour en faire le modèle à plusieurs niveaux ci-dessus de façon à obtenir le meilleur prédicteur linéaire sans biais (MPLSB) de la moyenne régionale $\mu_i = X_i' \beta_i$ en supposant N_i grand, où \bar{X}_i est le vecteur $p \times 1$ des moyennes connues de la population des variables auxiliaires pour la i -ième région. Ils ont également obtenu le MPLSB empirique et une approximation d'ordre deux de l'erreur quadratique moyenne (EQM) du MPLSB empirique pour le modèle à plusieurs niveaux. À l'aide de données sur le revenu des ménages de

- KNOWLES, J. (1997). Trend Estimation Practices of National Statistical Institutes. Office for National Statistics, Methods and Quality Division, UK, MQ 044.
- KNOWLES, J., et KENNY (1997). An Investigation of Trend Estimation Methods. Office for National Statistics, Methods and Quality Division, UK, MQ 044.
- LEE, H. (1990). Estimation des coefficients de corrélation de panel pour l'enquête sur la population active du Canada. *Techniques d'enquête*, 16, 297-306.
- LINACRE, S., et ZARB, J. (1991). Picking turning points in the economy. *Australian Economic Indicators*. Australian Bureau of Statistics, numéro 1350.0 au catalogue.
- MCLAREN, C.H. (1999). Designing Rotation Patterns and Filters for Trend Estimation in Repeated Surveys. Non publiée. Thèse de doctorat, School of Mathematics and Applied Statistics, University of Wollongong.
- MCLAREN, C.H., et STEEL, D.G. (1997). The effect of different rotation patterns on the sampling variance of seasonal and trend filters. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1997, 790-795.
- MIYAZAKI, E.S., et DOREA, C.C.Y. (1993). Estimation of the parameters of a time series subject to the error of rotation sampling. *Communications in Statistics*, A, 22, 805-825.
- PFEBBERMANN, D. (1994). A general method for estimating the variances of X-11 seasonally adjusted estimators. *Journal of Time Series Analysis*, 15, 85-116.
- RAO, J.N.K., et GRAHAM, J.E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 69, 492-509.
- SCOTT, A.J., SMITH T.M.F., et JONES, R. G. (1977). The application of time series methods to the analysis of repeated surveys. *Revue Internationale de Statistique*, 45, 3-73.
- SHISKIN, J., YOUNG, A.H., et MUSGRAVE, J.C. (1967). *X-11 Variant of the Census Method II Seasonal Adjustment Program*. Article technique 15, Bureau of the Census, U.S. Department of Commerce, Washington, D.C.
- YOUNG, A.H. (1968). Linear Approximations to the Census and BLS Seasonal Adjustment Methods. *Journal of the American Statistical Association*, 63.
- SINGH, M.P., DREW, J.D., GAMBINO, J., et MAYDA, F. (1990). Méthodologie de l'enquête sur la population active du Canada. Numéro 71-526 au catalogue, Statistique Canada.
- SMITH, T.M.F. (1997). Discussion of paper by Steel. *Journal of the Royal Statistical Society A*, 160, 33-34.
- STEEL, D.G. (1996). Options for Producing Monthly Estimates of Unemployment According to the ILO Definition. Central Statistical Office, U.K.
- STEEL, D.G. (1997). Producing monthly estimates of unemployment and employment according to the international labour office definition. *Journal of the Royal Statistical Society A*, 160, 5-46.
- STEEL, D.G., et MCLAREN, C.H. (2000). The effect of different rotation patterns on the revisions of trend estimates. *Journal of Official Statistics*, 16, 61-76.
- STEEL, D.G., et DEMEL, R. (1988). The Contribution of Sampling Error to the Variability of Statistical Series. Article présenté à la National Mathematical Sciences Congress, Canberra.
- STEEL, D.G. (1993). *X-11 Time Series Decomposition and Sampling Errors*. Document de travail dans Econometrics and Applied Statistics, numéro 93/2. Australian Bureau of Statistics, numéro 1351 au catalogue.
- STEEL, D.G. (1995). Seasonal Analysis and Sample Design. Article présenté à The Conference of Survey Measurement and Process Quality. Bristol 1995.
- TILLER, R.B. (1992). Time series modeling of sample survey data from the U.S. Current Population Survey. *Journal of Official Statistics*, 8, 149-166.
- WALLIS, K.F. (1982). Seasonal adjustment and revision of current data: linear filters for the x11-method. *Journal of the Royal Statistical Society A*, 145, 74-85.
- WOLTER, K.M., et MONSOUR, N.J. (1981). On the problem of variance estimation for a deseasonalized series. *Current Topics in Survey Sampling*, (D. Krewski, R. Platek et J.N.K. Rao, Eds.), Academic Press, New York, 367-407.

montre que les plans de renouvellement 1-2-(m) conviennent bien si l'évaluation de la tendance consiste à examiner les variations d'estimations désaisonnalisées sur trois ou sur six mois. Ces résultats donnent aussi à penser que ce genre de plan de renouvellement donne de bons résultats pour les estimations de la variation des estimations de la tendance sur les périodes de trois ou de six mois les plus récentes. Le critère d'évaluation utilisé ici est la variance d'échantillonnage des estimations de la tendance et des estimations désaisonnalisées, facteur qui est influencé par le plan d'échantillonnage. Steel et McLaren (2000) évaluent divers plans de renouvellement en ce qui concerne l'importance des corrections de ces estimations aux extrémités de la série et tirent des conclusions comparables quant aux plans de renouvellement.

REMERCIEMENTS

Les présents travaux ont été financés par l'*Australian Research Council* et l'*Australian Bureau of Statistics (ABS)*. Les opinions exprimées dans le présent article ne représentent pas nécessairement celles de ces organismes. Nous remercions le rédacteur adjoint et les examinateurs de leurs commentaires, ainsi que Geoff Lee, Andrew Sutcliffe et Phillip Bell de l'ABS, et Norma Chhab, de Statistique Canada.

BIBLIOGRAPHIE

- AUSTRALIAN BUREAU OF STATISTICS (1993). *A Guide to Interpreting Time Series – Monitoring "Trends"*. An Overview. Australian Bureau of Statistics, numéro 1348.0 au catalogue, Canberra.
- AUSTRALIAN BUREAU OF STATISTICS (1992). *Information Paper: Labour Force Survey Sample Design*. Australian Bureau of Statistics, numéro 6269.0 au catalogue, Canberra.
- AUSTRALIAN BUREAU OF STATISTICS (1987). *A Guide to Smoothing Time Series – Estimates of "Trend"*. Australian Bureau of Statistics, numéro 1316.0 au catalogue, Canberra.
- BELL, P.A. (1998). *Using State Space Models and Composite Estimation to Measure the Effects of Telephone Interviewing on Labour Force Estimates*. Document de travail dans *Econometrics and Applied Statistics*, 98/2, Australian Bureau of Statistics, numéro 1351.0 au catalogue, Canberra.
- BELL, P.A. (1999). *The Impact of Sample Rotation Patterns and Composite Estimation on Survey Outcomes*. Document de travail dans *Econometrics and Applied Statistics*, No. 99/1, Australian Bureau of Statistics, numéro 1352.0 au catalogue, Canberra.
- BELL, W.R., et HILLMER, S. (1990). Estimation dans les enquêtes à passages répétés du moyen de séries chronologiques. *Techniques d'enquête*, 16, 205-227.
- BELL, W.R., et KRAMER, M. (1999). Vers des variances pour la désaisonnalisation X-11. *Techniques d'enquête*, 25, 13-32.
- KISH, L. (1998). Space/time variations and rolling samples. *Journal of the Royal Statistical Society A*, 145, 1-41.
- KENNY, P.B., et DURBIN, J. (1982). Local trend estimation and seasonal adjustment of economic and social time series. *Journal of the Royal Statistical Society A*, 145, 1-41.
- KALTON, G., et CITRO, C.F. (1993). Enquêtes par panel: ajout d'une quatrième dimension. *Techniques d'enquête*, 19, 217-227.
- HAUSMAN, J.A., et WATSON, M.W. (1985). Error in variables and seasonal adjustment procedures. *Journal of the American Statistical Association*, 80, 531-540.
- HENDERSON, R. (1916). Note on graduation by adjusted averages. *Transactions of the Actuarial Society of America*, 17, 43-48.
- GRAY, A., et THOMSON, P. (1996). Design of moving-average trend filters using fidelity, smoothness and minimum revisions criteria. *Time Series Analysis in Memory of E.J. Hannan*. (P. Robinson et M. Rosenblatt, Eds.), 205-219. Springer lecture notes in statistics, 115.
- GRAY, A., et THOMSON, P. (1996). Design of moving-average trend filters using fidelity, smoothness and minimum revisions criteria. *Time Series Analysis in Memory of E.J. Hannan*. (P. Robinson et M. Rosenblatt, Eds.), 205-219. Springer lecture notes in statistics, 115.
- HENDERSON, R. (1916). Note on graduation by adjusted averages. *Transactions of the Actuarial Society of America*, 17, 43-48.
- HAUSMAN, J.A., et WATSON, M.W. (1985). Error in variables and seasonal adjustment procedures. *Journal of the American Statistical Association*, 80, 531-540.
- KALTON, G., et CITRO, C.F. (1993). Enquêtes par panel: ajout d'une quatrième dimension. *Techniques d'enquête*, 19, 217-227.
- KENNY, P.B., et DURBIN, J. (1982). Local trend estimation and seasonal adjustment of economic and social time series. *Journal of the Royal Statistical Society A*, 145, 1-41.
- KISH, L. (1998). Space/time variations and rolling samples. *Journal of Official Statistics*, 14, 31-46.
- McLaren et Steel: L'effet de divers plans de renouvellement sur la variance d'échantillonnage

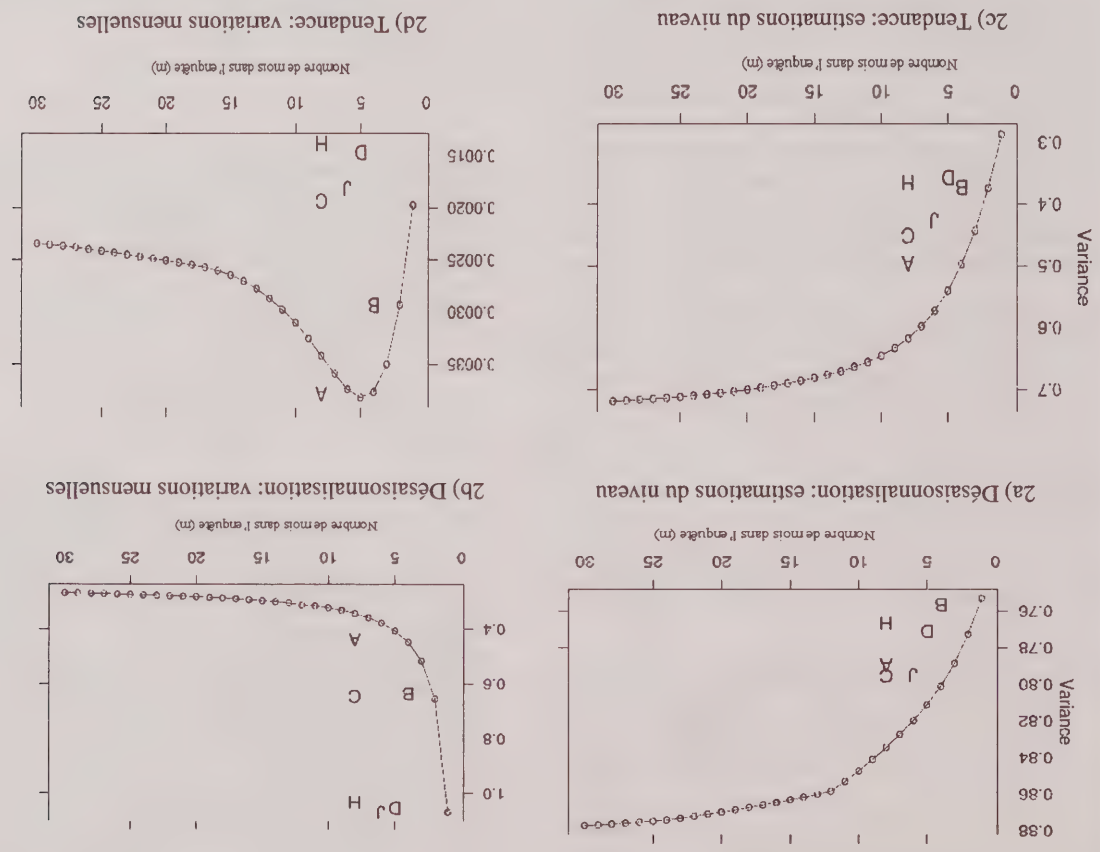
Les plans de renouvellement utilisés à l'heure actuelle, tels que dans-pour-8, dans-pour-6 et 4-8-4(8), sont raisonnables si la statistique principale sur laquelle l'analyse est la variation mensuelle des estimations désaisonnalisées. Selon nous, très souvent, l'examen de cette variation n'est pas un moyen fiable d'évaluer les tendances courantes. Il est nécessaire d'étudier la courbe de variation au cours des mois récents. On peut, pour cela, utiliser des filtres qui produisent une estimation de la tendance. Les résultats obtenus ici donnent à penser que, si l'utilisation des données d'enquête visent principalement à déterminer la tendance, il est préférable d'utiliser des plans de renouvellement assez différents. Plus précisément, les plans de renouvellement 1-2-1(m) donnent de bons résultats si l'on cherche à réduire la variance des estimations du niveau de la tendance et de l'écart entre deux estimations consécutives de la tendance pour une gamme de combinaisons de filtres. Les plans de renouvellement 1-2-1(m) donnent aussi de

6. DISCUSSION

bons résultats pour la variance d'échantillonnage des estimations désaisonnalisées de niveau. Par conséquent, lors de l'élaboration du plan de renouvellement de l'échantillon d'une enquête répétée, il faut tenir compte de l'importance relative des estimations désaisonnalisées et des estimations de la tendance. L'examen des figures 1 et 2 montre que le plan de renouvellement 2-2-2(8) est un compromis raisonnable si l'on juge important d'étudier à la fois le niveau et la variation mensuelle des estimations désaisonnalisées et des estimations de la tendance. Bell (1999) considère aussi l'effet de quatre plans de renouvellement distincts sur la variance d'échantillonnage du niveau et de la variation mensuelle des estimations originales, non corrigées, et des estimations de la tendance, l'évaluation de cette dernière demande l'examen de la variation d'estimations désaison-

Même si les analystes n'utilisent pas officiellement les plans de renouvellement 2-2-2(8) est un bon compromis. rhode X11 et d'une MMH sur 13 termes. Il juge aussi que

Figure 2. Rapport de la variance d'échantillonnage à la variance de la série originale pour le plan de renouvellement choisi pour la combinaison 4 (X11 ARIMA) pour la variable "personne occupée" où A=4-8-4(8), B=2-10-2(4), C=2-2-2(8), D=1-2-1(5), H=1-2-1(8), J=1-1-1(6).



respectivement. Les figures 2(a) à 2(d) présentent les résultats pour la combinaison 4 dans le cas de la variable "personne occupée".

Les colonnes 1 et 2 des tableaux 4, 5 et 6 montrent que les plans de renouvellement avec faible chevauchement mensuel donnent des résultats presque aussi bons que les plans à renouvellement complet pour les estimations déséquilibrées du niveau. La variance est plus forte pour les plans de renouvellement avec chevauchement mensuel important, particulièrement pour la combinaison 3 qui correspond à l'utilisation de la MMH sur 9 termes.

L'écart entre les rapports calculés pour les quatre combinaisons est minime dans le cas des estimations déséquilibrées de la variation mensuelle (colonnes 3 et 4 dans tous les tableaux). Les plans de renouvellement avec chevauchement mensuel important demeurent supérieurs aux plans avec chevauchement mensuel faible ou nul, quelle que soit la combinaison X11/X12ARIMA utilisée.

En ce qui concerne les estimations du niveau de la tendance, les plans de renouvellement avec chevauchement important de l'échantillon produisent de nouveau un rapport plus grand des variances. Les plans de renouvellement 1-2-1(5) et 1-2-1(8) continuent de donner de meilleurs résultats que les autres pour chaque combinaison de filtres, mais donnent de moins bons résultats qu'un échantillon indépendant pour les combinaisons 2 et 4.

Dans le cas des estimations de la variation mensuelle de la tendance, les meilleurs plans de renouvellement sont de nouveau les plans 1-2-1(5) et 1-2-1(8), qui donnent des résultats supérieurs à ceux de l'échantillon indépendant pour les quatre combinaisons de filtres. Pour la combinaison 3, les plans de renouvellement avec chevauchement mensuel important sont aussi bons que les plans de renouvellement 1-2-1(m). Pour les combinaisons 2 et 3, le plan 1-1-1(6) est légèrement supérieur au plan 1-2-1(m). Le remplacement des plans de renouvellement utilisés à l'heure actuelle par les plans de renouvellement 1-2-1(m) produirait des améliorations considérables. Par exemple, pour la variable d'emploi, passer du plan 4-8-4(8) au plan 1-2-1(8) produirait des gains de 42%, 25% et 64% en choisissant les combinaisons 2, 3 et 4, respectivement.

Ces résultats se fondent sur les estimations de corrélation de l'BAPA qui, puisqu'elles sont fondées sur des estimations d'enquête, sont entachées d'une erreur d'échantillonnage. Les filtres de la tendance étudiés ne sont pas calculés en se servant de ces estimations. Nous arrivons aux mêmes conclusions générales quant à l'effet de divers plans de renouvellement pour les deux modèles de corrélation basés sur des corrélations raisonnablement différentes. Selon nous, les conclusions sont applicables à la gamme de modèles de corrélation compris entre ces deux modèles extrêmes. McLaren et Steel (1997) arrivent aux mêmes conclusions en se servant du modèle de corrélation calculé par Steel (1996) pour l'emploi et le chômage au Royaume-Uni.

En ce qui concerne les estimations du niveau de la tendance, la variance augmentée parallèlement au chevauchement des échantillons mensuels (voir la figure 1(c), ainsi que les colonnes 5 et 6 du tableau 3). Dans le cas des plans de renouvellement *dans-sus-pour-m*, la variance augmente rapidement quand *m* passe de 1 à 5. Les plans de renouvellement 1-2-1(5) et 1-2-1(8) donnent d'assez bons résultats que si l'on sélectionnait un échantillon indépendant chaque mois et de nettement meilleurs résultats que les plans de renouvellement avec chevauchement mensuel. Cette situation tient surtout au fait que, pour une moyenne mobile, il est préférable de calculer la moyenne d'observations indépendantes que celle d'observations positivement corrélées. La variance plus grande du plan 1-1-1(6) que des plans 1-2-1(5) et 1-2-1(8) donne à penser que, dans le cas de plans de renouvellement sans chevauchement mensuel, l'intervalle entre deux inclusions successives des unités d'échantillon exerce un certain effet.

La figure 1(d), ainsi que les colonnes 7 et 8 du tableau 3 montrent que, pour les estimations de la variation mensuelle de la tendance, la variance augmente très rapidement quand *m* passe de 1 à 3 et diminue très rapidement quand *m* augmente à partir de 4. De tous les plans de renouvellement considérés, le plan *dans-pour-3* semble être le moins bon et le plan utilisé à l'heure actuelle pourrait donc être amélioré considérablement. Par exemple, le choix d'un plan 1-2-1(8) au lieu de 4-8-4(8) réduirait la variance des estimations de la variation mensuelle de la tendance de 55% pour l'emploi et de 43% pour le chômage. Bien que le degré de chevauchement mensuel demeure un facteur essentiel, le plan d'inclusion joue un rôle équilibrant. Par exemple, le plan 2-2-2(8) produit une variance plus faible que le plan *dans-pour-2* ou 2-10-2(4). Qui plus est, pour les estimations de la variation mensuelle de la tendance, les plans de renouvellement les meilleurs sont les plans 1-2-1(5) et 1-2-1(8), qui donnent des résultats nettement supérieurs à ceux obtenus en procédant à un renouvellement complet chaque mois. Ce résultat tient au fait que les estimations de la variation mensuelle de la tendance représentent, en réalité, l'écart entre des séries désaisonnalisées décalées de quelques mois et que les plans de renouvellement 1-2-1(m) produisent des corrélations positives entre les estimations espacées de trois mois. McLaren et Steel (1997) ont obtenu des résultats comparables en se servant de l'approximation de X11 proposée par Sutcliffe (1993).

Les résultats montrent que, pour les estimations du niveau et du dernier mouvement de la tendance, les plans de renouvellement 1-2-1(m) produisent une variance d'échantillonnage nettement plus faible que les plans de renouvellement utilisés à l'heure actuelle.

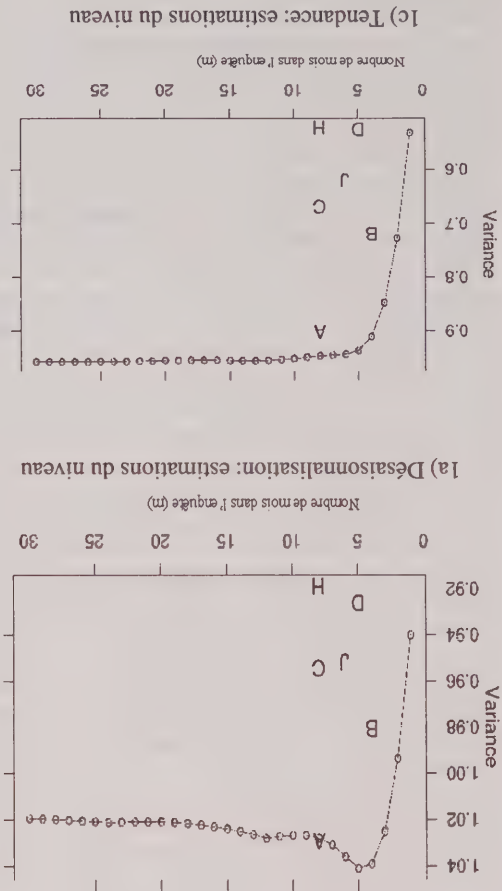
5.2 X12ARIMA – Filtres en cascade concourants avec extrapolation

Les résultats de l'application des combinaisons de filtres 2, 3 et 4 sont présentés aux tableaux 4, 5 et 6,

5.1 X11 – Filtrés en cascade types concourants

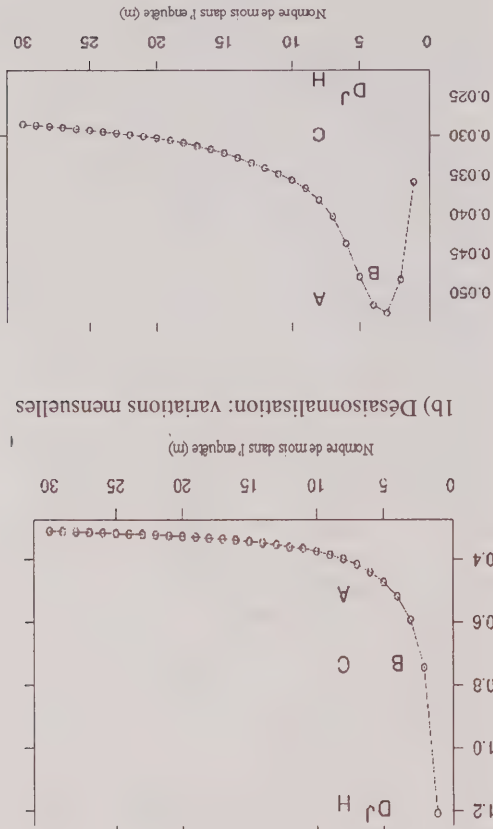
Les résultats obtenus en utilisant les filtres types X11 (combinaison 1) sont présentés au tableau 3. Les figures 1(a) à 1(d) montrent la variance d'échantillonnage du niveau et de la variation mensuelle des estimations désaisonnalisées et de la tendance à la fin de la série divisée par la variance de l'estimation originale du niveau en fonction du nombre total de fois qu'une unité sélectionnée est incluse. Les résultats pour la variable employée sont portés en graphique pour certains plans *a-b-a(m)* et pour le plan de renouvellement *dans-pour-m* où *m* varie de 1 à 30. Un plan de renouvellement *dans-pour-30* signifie qu'il n'y a pas de renouvellement.

Les colonnes 1 et 2 du tableau 3 montrent que les plans de renouvellement sans chevauchement mensuel donnent de bons résultats en ce qui concerne la variance des estimations désaisonnalisées du niveau. L'utilisation de plans de



1(a) Désaisonnalisation: estimations du niveau

1(c) Tendance: estimations du niveau



1(b) Désaisonnalisation: variations mensuelles

1(d) Tendance: variations mensuelles

Figure 1. Rapport de la variance d'échantillonnage à la variance de la série originale pour le plan de renouvellement choisi pour la combinaison 1 (X11) pour la variable personne occupée où A = 4-8-4(8), B = 2-10-2(4), C = 2-2-2(8), D = 1-2-1(5), H = 1-2-1(8), J = 1-1-1(6).

observée sans erreur d'échantillonnage, ne dépend pas du plan d'échantillonnage, y compris le plan de renouvellement. Bell et Kramer (1999) ont étudié la série sur les mises en chantier aux États-Unis comportant au moins cinq unités et montrent que la variance totale de la série augmente fortement aux extrémités à cause d'erreurs de prévision. Cette situation est due aux corrections des estimations initiales de la tendance faites à mesure que l'on ajoute des estimations à la série. Steel et McLaren (2000) étudient l'effet de divers plans de renouvellement sur la correction observée des estimations initiales de la tendance, que l'on peut représenter par $w_{t,T}^* y_{t,T}^* - w_{t,T}^* y_{t,T}^*$. Ils constatent que l'importance relative de la composante due à l'erreur d'échantillonnage dépend de la façon dont la série réelle évolue autour de la période observée.

5. RÉSULTATS

Nous utilisons les filtres qui correspondent au niveau, d'une part, et à la variation mensuelle, d'autre part, pour les estimations désaisonnalisées ainsi que les estimations de la tendance à la toute extrémité de la série. Les tableaux 3 à 6 résument les effets de divers plans de renouvellement pour chaque combinaison de filtres en cascade. Ces tableaux donnent, pour certains plans de renouvellement, le rapport de la variance d'échantillonnage que l'on obtiendrait si l'on renouvelait complètement l'échantillon chaque mois. Les rapports calculés pour le milieu de la série mènent à des conclusions générales semblables (McLaren 1999).

Tableau 3

Rapport de la variance d'échantillonnage pour certains plans de renouvellement à la variance d'échantillonnage indépendant d'échantillonnage (Combinaison 1)

Plan de renouvellement	S_A'	$S_{A+1}' - S_A'$	T_t'	$T_{t+1}' - T_t'$
complet	1,00	1,00	1,00	1,00
1-2-(1(5))	0,99	0,99	1,00	0,68
1-2-(1(8))	0,98	0,99	0,97	0,64
1-1-(1(6))	1,01	1,01	1,00	0,70
2-2-(2(8))	1,02	1,02	0,61	0,71
2-10-(2(4))	1,04	1,04	0,61	0,71
3-3-(3(6))	1,07	1,06	0,48	0,61
4-8-(4(8))	1,10	1,08	0,42	0,57
6-6-(6(12))	1,10	1,08	0,36	0,52
dans-pour-6	1,10	1,08	0,36	0,52
dans-pour-8	1,09	1,08	0,33	0,50
pas de renouvellement	1,08	1,08	0,24	0,44

Rapport de la variance d'échantillonnage pour certains plans de renouvellement à la variance d'échantillonnage

(Combinaison 4)

Plan de renouvellement	S_A'	$S_{A+1}' - S_A'$	T_t'	$T_{t+1}' - T_t'$
complet	1,00	1,00	1,00	1,00
1-2-(1(5))	1,02	1,02	0,99	1,19
1-2-(1(8))	1,02	1,02	0,97	1,21
1-1-(1(6))	1,06	1,04	1,00	1,39
2-2-(2(8))	1,06	1,04	0,60	0,71
2-10-(2(4))	1,00	1,01	0,60	0,70
3-3-(3(6))	1,07	1,05	0,48	0,61
4-8-(4(8))	1,05	1,04	0,41	0,56
6-6-(6(12))	1,08	1,06	0,35	0,51
dans-pour-6	1,09	1,07	0,35	0,52
dans-pour-8	1,11	1,08	0,32	0,49
pas de renouvellement	1,17	1,12	0,24	0,43

Tableau 6

Rapport de la variance d'échantillonnage pour certains plans de renouvellement à la variance d'échantillonnage indépendant d'échantillonnage (Combinaison 4)

Plan de renouvellement	S_A'	$S_{A+1}' - S_A'$	T_t'	$T_{t+1}' - T_t'$
complet	1,00	1,00	1,00	1,00
1-2-(1(5))	0,99	0,99	0,96	0,98
1-2-(1(8))	0,97	0,99	0,93	0,97
1-1-(1(6))	1,04	1,02	0,99	1,08
2-2-(2(8))	1,07	1,06	0,60	0,71
2-10-(2(4))	1,05	1,06	0,61	0,72
3-3-(3(6))	1,15	1,12	0,51	0,63
4-8-(4(8))	1,12	1,11	0,44	0,58
6-6-(6(12))	1,14	1,13	0,37	0,53
dans-pour-6	1,16	1,13	0,38	0,53
dans-pour-8	1,17	1,14	0,34	0,51
pas de renouvellement	1,22	1,17	0,25	0,44

Tableau 5

Rapport de la variance d'échantillonnage pour certains plans de renouvellement à la variance d'échantillonnage indépendant d'échantillonnage (Combinaison 3)

Plan de renouvellement	S_A'	$S_{A+1}' - S_A'$	T_t'	$T_{t+1}' - T_t'$
complet	1,00	1,00	1,00	1,00
1-2-(1(5))	0,99	0,99	0,96	0,98
1-2-(1(8))	0,97	0,99	0,93	0,97
1-1-(1(6))	1,04	1,02	0,99	1,08
2-2-(2(8))	1,07	1,06	0,60	0,71
2-10-(2(4))	1,05	1,06	0,61	0,72
3-3-(3(6))	1,15	1,12	0,51	0,63
4-8-(4(8))	1,12	1,11	0,44	0,58
6-6-(6(12))	1,14	1,13	0,37	0,53
dans-pour-6	1,16	1,13	0,38	0,53
dans-pour-8	1,17	1,14	0,34	0,51
pas de renouvellement	1,22	1,17	0,25	0,44

Tableau 4

Rapport de la variance d'échantillonnage pour certains plans de renouvellement à la variance d'échantillonnage indépendant d'échantillonnage (Combinaison 2)

Plan de renouvellement	S_A'	$S_{A+1}' - S_A'$	T_t'	$T_{t+1}' - T_t'$
complet	1,00	1,00	1,00	1,00
1-2-(1(5))	1,01	1,01	0,99	1,05
1-2-(1(8))	1,00	1,00	0,96	0,99
1-1-(1(6))	1,04	1,03	1,00	1,22
2-2-(2(8))	1,05	1,04	0,60	0,71
2-10-(2(4))	1,02	1,03	0,60	0,71
3-3-(3(6))	1,08	1,06	0,49	0,61
4-8-(4(8))	1,06	1,06	0,41	0,56
6-6-(6(12))	1,08	1,07	0,35	0,52
dans-pour-6	1,10	1,08	0,36	0,52
dans-pour-8	1,11	1,08	0,32	0,49
pas de renouvellement	1,14	1,11	0,24	0,43

approche pour obtenir une approximation réaliste de ces deux dernières méthodes.

Représentons par H_{13} la matrice dont les lignes contiennent les poids de filtrage des moyennes mobiles de

Henderson sur 13 termes pour les filtres symétriques ainsi qu'asymétriques. Représentons par $S_{3 \times 3}$ la matrice des poids qui correspondent à la moyenne mobile (mm) 3×3 et par $S_{3 \times 5}$, celle qui correspond à la mm 3×5 . Nous

utilisons ces matrices pour estimer les facteurs de saisonnalité. Enfin, nous représentons par D une mm centrée sur 12 termes et par I , une matrice d'identité. La notation c

indique le complément d'un filtre, par exemple $D^c = I - D$. Les filtres de désaisonnalisation en cascade sont représentés

$$S = I - D^c S_{3 \times 5} [H_{13} (D^c S_{3 \times 3} D^c)^c];$$

Nous obtenons alors le filtre en cascade pour estimer la

tendance en multipliant le filtre de désaisonnalisé par un filtre

de la tendance. Aux extrémités de la série, les filtres en cascade appliqués aux estimations de la tendance et aux estimations désaisonnalisées diffèrent selon qu'on applique la

méthode X_{11} ou $X_{11}ARIMA$.

Nous considérons les combinaisons des filtres internes de X_{11} et de $X_{11}ARIMA$ qui suivent:

1. Filtre en cascade X_{11} type: Ce filtre comprend une MMH sur 13 termes pour l'estimation de la tendance (H_{13}), une mm 3×3 pour la première estimation des facteurs saisonniers ($S_{3 \times 3}^1$), une mm 3×5 pour l'estimation des facteurs saisonniers ($S_{3 \times 5}^2$), mais ne comporte aucune modification pour les valeurs aberrantes.
2. Filtre en cascade X_{11} type avec prévisions ARIMA: Ce filtre comprend les filtres H_{13} , $S_{3 \times 3}^1$ et $S_{3 \times 5}^2$, ainsi que des prévisions extrapolées d'après un modèle ARIMA de la forme $(1 - B)(1 - B^{12})Y_t = (1 - 0.4B)(1 - 0.6B^{12})a_t$, où B est l'opérateur de rétrogradation et a_t , un facteur de traitement de l'erreur résiduelle (bruit blanc), mais ne comporte aucune modification pour les valeurs aberrantes.
3. Filtre en cascade X_{11} court avec prévisions ARIMA: Ce filtre comprend les filtres H_9 , $S_{3 \times 3}^1$ et $S_{3 \times 5}^2$, ainsi que des prévisions extrapolées d'après un modèle de la forme $(1 - B)(1 - B^{12})Y_t = (1 - 0.3B)(1 - 0.3B^{12})a_t$, mais ne comporte aucune modification pour les valeurs aberrantes.
4. Filtre en cascade X_{11} long avec prévisions ARIMA: Ce filtre comprend les filtres H_{23} , $S_{3 \times 3}^1$ et $S_{3 \times 5}^2$, ainsi que des prévisions extrapolées d'après un modèle de la forme $(1 - B)(1 - B^{12})Y_t = (1 - 0.8B)(1 - 0.8B^{12})a_t$, mais ne comporte aucune modification pour les valeurs aberrantes.

Selon Dagum (1983), les combinaisons 2 et 3 sont applicables à plusieurs cas. Les approximations linéaires choisies nous permettent d'examiner l'effet de divers plans

de renouvellement pour une gamme de filtres axés sur des MMH de diverses longueurs que l'on utilise en pratique. Pour chaque combinaison de filtres, le filtre en cascade correspondant fournit un vecteur de poids de filtrage pour l'estimation finale de la tendance. Ces vecteurs peuvent être introduits dans l'équation (2) pour obtenir la variance d'échantillonnage pour un plan de renouvellement particulier grâce à l'utilisation des valeurs appropriées de $V(Y^T | X^T)$. Si l'on estime les variations, le vecteur de données Y^T reste le même, mais les poids appliqués changent. Par exemple, on peut utiliser $w_{t+1} - w_t$ pour un écart d'un mois. Cette méthode de base est la même que celle adoptée par Wolter et Monsour (1981) qui ont proposé d'estimer la variance des estimations désaisonnalisées en se servant de l'équation (2) avec des poids choisis de façon à obtenir une approximation raisonnable de la méthode de désaisonnalisation et d'utiliser une estimation de $V(Y^T | X^T)$ fondée sur des données d'enquête. Nous considérons aussi les filtres de la tendance et différentes réalisations de $X_{11}ARIMA$ et des plans de renouvellement.

Les modèles $X_{11}ARIMA$ étudiés ici sont représentatifs de ceux utilisés couramment en pratique. L'utilisation de prévisions ARIMA, dans le cadre de la méthode $X_{11}ARIMA$, complique la situation. Par exemple, on suppose que les spécifications du modèle ARIMA ne comportent aucune erreur. Habituellement, on identifie et on estime le modèle ARIMA d'après des données d'enquête antérieures. Or, l'erreur d'échantillonnage commise lors de périodes antérieures risque d'influencer le choix du modèle ARIMA et des filtres X_{11} . On pourrait tenir compte de cette situation en modifiant le terme de variance dans (2). Nous calculons les premières estimations désaisonnalisées et de la tendance pour le temps t en nous servant de la série chronologique d'estimations qui se termine au temps t , c'est-à-dire y_t , ce qui donne la valeur filtrée w_t^* . La valeur que nous obtenons si aucune erreur d'échantillonnage n'était commise est $w_t^* X_t$. L'erreur d'échantillonnage considérée ici est $w_t^* y_t - w_t^* X_t$. Comme nous ajoutons des estimations à la série, la valeur filtrée au temps t pourrait changer, mais à partir d'un certain point, $t + s$, nous ne noterons plus aucune variation appréciable. La valeur filtrée finale au temps t basée sur les estimations d'enquête peut s'écrire $w_t^* y_{t+s}^*$, pour un vecteur final symétrique de poids w_t^* . Pareillement, la valeur finale que nous obtenons si aucune erreur d'échantillonnage n'était commise serait $w_t^* X_{t+s}^*$. Bell et Krammer (1999) considèrent la différence $w_t^* y_{t+s}^* - w_t^* X_{t+s}^*$, qui inclut l'erreur de prévision. Nous pouvons décomposer cette différence comme suit:

$$w_t^* y_{t+s}^* - w_t^* X_{t+s}^* = (w_t^* y_{t+s}^* - w_t^* y_t^*) + (w_t^* y_t^* - w_t^* X_{t+s}^*).$$

Nous avons étudié la façon dont divers plans de renouvellement influent sur le premier terme de cette décomposition. Le deuxième terme, qui correspond à la série

les proportions de personnes occupées et de chômeurs tirées de l'enquête australienne sur la population active (EAPA), sont présentées au tableau 2. Pour les obtenir, on a traité les groupes de renouvellement de l'EAPA comme des répétitions et on a déterminé l'autocorrélation au niveau du groupe de renouvellement. On s'est servi d'un modèle proposé par Bell (1998) pour extrapoler les valeurs au-delà des décalages donnés.

Autocorrélations – EAPA									
Tableau 2									
Proportion de personnes occupées					Proportion de chômeurs				
décalage	1	2	3	4	5	6	7	8	
$r(s)$	0,80	0,71	0,64	0,57	0,50	0,45	0,40	0,36	
$d(s)$	0,15	0,15	0,14	0,13	0,12	0,11	0,11	0,10	
1	0,11	0,11	0,10	0,09	0,09	0,08	0,08	0,07	
2	0,62	0,52	0,44	0,37	0,31	0,26	0,22	0,19	
3	0,11	0,11	0,10	0,09	0,09	0,08	0,08	0,07	
4	0,62	0,52	0,44	0,37	0,31	0,26	0,22	0,19	
5	0,11	0,11	0,10	0,09	0,09	0,08	0,08	0,07	
6	0,62	0,52	0,44	0,37	0,31	0,26	0,22	0,19	
7	0,11	0,11	0,10	0,09	0,09	0,08	0,08	0,07	
8	0,62	0,52	0,44	0,37	0,31	0,26	0,22	0,19	

Sutcliffe et Lee (1995) étudient les erreurs types des estimations désaisonnalisées et des estimations de la tendance des niveaux et des flux pour un petit nombre de plans de renouvellement distincts. Ils supposent pour cela que les corrélations entre les estimations d'enquête obéissent à un modèle de décroissance géométrique simple, avec, pour la population, une corrélation $\rho = 0,8$, c'est-à-dire $R(s) = \rho^s$, qui diminue plus rapidement que les valeurs données au tableau 2.

4. APPROXIMATIONS LINÉAIRES DES ESTIMATIONS DÉSAISONNALISÉES ET DES ESTIMATIONS DE LA TENDANCE

La méthode X11 consiste à appliquer itérativement des moyennes mobiles qui produisent un filtre symétrique pour les valeurs centrales et des filtres asymétriques pour les valeurs de début et de fin de série. Le lissage au moyen de filtres linéaires donne une approximation des estimations finales désaisonnalisées et des estimations finales de la tendance produites par la méthode X11. Plusieurs auteurs, dont Young (1968), Cleaveland et Tiao (1976), Wallis (1982) et Sutcliffe (1993), ont produit des approximations linéaires de la méthode X11. La méthode X11ARIMA (Dagum 1980, 1988), quant à elle, est une extension de la méthode X11 qui consiste à extrapoler la série originale aux deux extrêmes par application d'un modèle autorégressif à moyennes mobiles intégré (ARIMA pour *Autoregressive Integrated Moving Average Model*). On peut intégrer l'effet de l'extrapolation ARIMA à la pondération des filtres et puis appliquer les poids ainsi obtenus aux données uniquement. Dagum, Chhab et Chiu (1996) ont proposé une méthode en cascade, où les filtres en cascade résultent de la convolution de divers filtres linéaires prédéterminés utilisés dans les méthodes X11 et X11ARIMA. Nous utilisons leur

$$R(s) = d(s) + k(s)r(s) - d(s)) \tag{3}$$

renouvellement n'a eu lieu et $d(s)$ s'il y a eu renouvellement. Nous supposons aussi que l'estimation au temps t est égale, du moins approximativement, à la moyenne des estimations pour chaque groupe de renouvellement et que les estimations obtenues pour divers groupes de renouvellement, qui correspondent habituellement à des UPB différentes et bien séparées spatialement, sont indépendantes. Ces hypothèses sous-entendent que la corrélation d'échantillonnage entre y'_t et y'_{t+s} est

où $k(s)$ représente la proportion de l'échantillon en commun entre les deux périodes de référence. Le plan de renouvellement détermine le facteur $k(s)$ de chevauchement des échantillons. Par exemple, pour un plan de renouvellement *dans-pour-m*, $k(s) = 1 - s/m$, $s = 0, \dots, m - 1$ et est nul autrement, si l'on suppose que le même nombre de logements sont ajoutés et supprimés de l'échantillon chaque mois. Si les divers panels qui font partie d'un groupe particulier de renouvellement sont indépendants, alors $d(s) = 0$, mais, en général, cela n'est pas le cas. Ce modèle est essentiellement le même que celui dérivé par Scott, Smith et Jones (1977). Le tableau 1 donne un exemple de plan de renouvellement *dans-pour-4* appliqué sur une période de huit mois. Les divers panels sont représentés par des lettres différentes et l'indice indique le nombre de fois que le panel a été inclus dans l'échantillon d'enquête jusqu'à la période de référence indiquée.

Structure du plan de renouvellement dans-pour-4									
Tableau 1									
Période de référence					Groupe de renouvellement				
1	a_1	a_2	a_3	a_4	b_1	b_2	b_3	b_4	
2	c_1	d_1	d_2	d_3	d_4	e_1	e_2	e_3	
3	f_1	f_2	f_3	f_4	g_1	g_2	g_3	g_4	h_1
4	i_1	i_2	i_3	i_4	j_1	j_2	j_3	j_4	k_1

Ici, $r(2)$ est la corrélation provenant de, disons, a_2 et a_4 , tandis que $d(2)$ est la corrélation associée à a_2 et b_1 . Binder et Hidiroglou (1988), ainsi que Fuller et coll. (1992) discutent de la structure de données que suggèrent d'autres plans de renouvellement. L'hypothèse selon laquelle la variance de la série d'erreurs d'échantillonnage est constante sous-entend qu'une modification importante du plan d'échantillonnage n'affecte pas la structure de la population. Envisager l'hypothèse d'autocorrélations stables, $r(s)$ et $d(s)$, pour la corrélation au niveau de la population ne varie beaucoup. Les estimations de $r(s)$ et $d(s)$ dans (3) sont tirées d'une étude réalisée par Bell (1998). Les valeurs utilisées, c'est-à-dire

plan de renouvellement utilisé au Canada (Singh, Drew, Gambino et Mayda 1990) et celui où $m = 8$, à celui appliqué en Australie (ABS 1992). Steel (1997) a observé que l'enquête trimestrielle britannique sur la population active correspond à peu près à une enquête mensuelle avec un plan de renouvellement 1-2-1(5).

Nous considérons ici la variance d'échantillonnage des estimations désaisonnalisées et des estimations de la tendance associée aux plans de renouvellement utilisés à l'heure actuelle pour les EMPA et à un certain nombre de plans de renouvellement qui, même s'ils ne sont pas utilisés à l'heure actuelle, pourraient présenter certaines propriétés intéressantes. Cette étude donnera une idée des plans de renouvellement qui donnent les meilleurs résultats en ce qui concerne la composante de la variabilité des séries estimées sur laquelle le plan d'échantillonnage exerce une influence.

3. VARIANCE D'ÉCHANTILLONNAGE DES ESTIMATIONS DÉSAISONNALISÉES ET DES ESTIMATIONS DE LA TENDANCE

Représentons par y_T le vecteur qui contient les valeurs de la série chronologique des estimations d'enquête jusqu'au temps T et par X_T , le vecteur qui contient les valeurs réelles de population. La variance d'échantillonnage de la série originale est représentée par $V(y_T | X_T)$. Considérons un filtre linéaire utilisé pour obtenir des valeurs à partir de y_T par application d'un vecteur de poids de filtrage w_T . Les poids de filtrage ne sont pas aléatoires et ne présentent aucun lien avec les poids de sondage utilisés pour calculer les estimations d'enquête y_T . Le vecteur de poids de filtrage w_T dépend de la période de référence à laquelle se rapportent les valeurs filtrées. Les poids sont constants dans le corps de la série, mais peuvent être modifiés au début et à la fin de celle-ci. La valeur filtrée au temps t est donnée par

$$y_t = w'_t y_T \quad (1)$$

Alors

$$V(y_t | X_T) = w'_t V(y_T | X_T) w_t \quad (2)$$

représente la variance d'échantillonnage de la valeur filtrée au temps t . L'erreur d'échantillonnage de la valeur filtrée est égale à la différence entre $w'_t y_T$ et $w'_t X_T$, qui est subordonnée aux valeurs de la série réelle, X_T . Il s'agit de la différence entre la valeur filtrée obtenue d'après la série d'estimations qui se termine au temps T et la valeur que l'on obtiendrait si l'on observait cette série sans erreur d'échantillonnage. Nous nous concentrons sur cette composante, puisque c'est sur la variance d'échantillonnage que peut influencer la modification du plan d'échantillonnage. On ne tient pas compte de la variance associée à X_T . Wolter et Monsour (1981) ont examiné la question de la variance totale par opposition à la variance de l'erreur d'échantillonnage. Tenir compte de la variance totale pour interpréter la

L'analyse de l'effet de divers plans de renouvellement se simplifie si la structure d'autocorrélation de la série d'erreurs d'échantillonnage est stable. La forme précise de la fonction d'autocorrélation dépend de la série et doit refléter les complexités du plan d'échantillonnage. Par exemple, Steel et DeMel (1988) proposent un modèle pour les données de l'Enquête mensuelle australienne sur la population active, tandis que Bell et Wilcox (1993) en proposent un pour la série de données sur le commerce de détail aux États-Unis. Bell et Hillmer (1990), ainsi que Mizaki et Dorea (1993) considèrent aussi de modéliser les erreurs d'enquête au moyen de modèles de série chronologique. Dempster et Hwang (1993) et Lee (1990) étudient diverses méthodes en vue d'estimer et de modéliser les corrélations des erreurs d'échantillonnage dans le cas de la CPS améri-

Nous posons que la variance de la série d'erreurs d'échantillonnage, e_t , est constante. Nous devons préciser le modèle de la corrélation entre les erreurs d'échantillonnage de y_t et y_{t+s} . Tous les plans de renouvellement considérés sous-entendent que, à tout point dans le temps, l'échantillon comprendra plusieurs panels. Un panel est un ensemble d'unités qui sont intégrées dans l'enquête et qui sont éliminées en même temps. Tout panel retiré de l'échantillon est remplacé par un autre. L'ensemble de panels reliés de cette façon forme un groupe de renouvellement. La plupart des EMPA s'appuient sur l'échantillonnage à plusieurs degrés et tout panel qui est éliminé de l'enquête est remplacé par un autre panel de ménages voisins (voir ABS 1992, Singh et coll. 1990). Par conséquent, nous supposons que la corrélation d'échantillonnage entre les estimations obtenues auprès d'un même groupe de renouvellement à s périodes d'intervalle est $r(s)$ si aucun

explicitement la série chronologique. D'autres auteurs, comme Bell et Wilcox (1993), Tiller (1992), Burridge et Wallis (1985) et Hausman et Watson (1985), étudient l'application de modèles ARIMA explicites aussi bien à la série réelle qu'à la série d'erreurs d'échantillonnage, et se concentrent sur l'estimation des paramètres des modèles. Ces travaux ne tiennent pas compte de l'effet des divers plans de renouvellement et visent essentiellement à produire des estimations de la variance des estimations désaisonnalisées pour le plan de renouvellement particulier utilisé.

Le plan de renouvellement de l'échantillon choisi pour l'enquête influence la structure d'autocorrélation de la série d'erreurs d'échantillonnage, donc la variance d'échantillonnage des estimations désaisonnalisées et des estimations originales de la tendance. Le choix du plan de renouvellement dépend de plusieurs facteurs. Un chevauchement important entre les échantillons de deux périodes consécutives réduit la variance d'échantillonnage des estimations de la variation d'une période à l'autre, tandis qu'un chevauchement important entre des périodes séparées par un intervalle de 12 mois réduit la variance d'échantillonnage des estimations de la variation annuelle. Habituellement, la première inclusion d'une unité sélectionnée dans l'enquête est celle qui est la plus coûteuse. Le fait de retenir les unités sélectionnées dans l'échantillon pendant une longue période réduit le coût de l'enquête. On établit donc des plans de renouvellement en vertu desquels une unité sélectionnée est incluse pendant une série aussi longue que possible de périodes consécutives. Cependant, toute unité sélectionnée doit finalement être éliminée de l'échantillon. Outre la nécessité de répartir le fardeau de réponse pour des raisons d'éthique, le fait de retenir une unité dans l'échantillon pendant un grand nombre de périodes d'observation risque de faire baisser le taux de réponse et la qualité des données recueillies (pour une discussion de ces problèmes, voir Kalton et Citro 1993).

Les plans de renouvellement varient en ce qui a trait au nombre de fois qu'une unité est incluse dans l'échantillon de l'enquête et à l'intervalle entre les inclusions. Nous nous concentrons ici sur les enquêtes mensuelles sur la population active (BMPA). En pratique, les plans de renouvellement adoptés sont des cas spéciaux du plan de renouvellement $a-b-a(m)$ selon lequel les unités sélectionnées sont incluses pendant a mois consécutifs, retirées de l'échantillon pendant b mois, puis réintroduites dans l'échantillon pour une période supplémentaire de a mois. Le plan est répété de sorte que les unités sélectionnées soient incluses, en tout, pour m occasions. Rao et Graham (1964) considèrent l'estimation des moyennes et des totaux de population finies pour sept classes de plans de renouvellement. Aux États-Unis, la *Current Population Survey (CPS)* est exécutée selon un plan 4-8-4(8) (Fuller, Adam et Yansaneh 1992). Si l'on pose $b = 0$, on obtient un plan de renouvellement *dans-pour-m* selon lequel les logements sélectionnés sont inclus pendant m mois après quoi ils sont éliminés de l'échantillon. Le cas où $m = 6$ correspond au

où e_t est l'erreur d'échantillonnage. On considère que la série Y_t comprend les composantes de la tendance-cyclo, de la saisonnalité et des jours irréguliers T_t , S_t et I_t , si bien que

$$y_t = T_t + S_t + I_t + e_t.$$

Dans certains cas, une décomposition multiplicative peut être plus appropriée. Nombre de bureaux de la statistique produisent des séries désaisonnalisées en essayant d'estimer S_t et de l'éliminer de la série, habituellement au moyen d'une combinaison de filtres linéaires. Les méthodes utilisées le plus couramment sont la méthode Census X11 mise au point par Shiskin et coll. (1967) et la variante X11ARIMA élaborée par Dagum (1980 et 1988). Findley, Monseil, Otto, Bell et Pugh (1998) ont décrit d'autres améliorations qui sont intégrées à la variante X12ARIMA. L'ABS publie aussi des estimations de la tendance obtenues par application des MMH à la série désaisonnalisée et encourage les utilisateurs à interpréter la série en se basant sur ces estimations de la tendance (Linacre et Zarb 1991; ABS 1993). Les moyennes mobiles de Henderson (1916), mises au point au départ pour des calculs actuariats, sont utilisées dans X11, X11ARIMA et X12ARIMA pour extraire la tendance des séries aux fins de la désaisonnalisation. Kenny et Durbin (1982) et Gray et Thomson (1996) expliquent le calcul des MMH. L'utilisateur des séries peut aussi produire des estimations de la tendance en appliquant les filtres aux estimations désaisonnalisées publiées. Kenny et Durbin (1982) font remarquer que la définition de la tendance n'est pas unique et que l'on peut choisir des filtres différents selon le degré de lissage et de sensibilité souhaitée. Knowles et Kenny (1997) ont étudié des méthodes d'estimations de la tendance en vue de produire des séries statistiques officielles. Pour les séries mensuelles, ils recommandent l'utilisation des MMH, en choisissant un filtre de 13 ou de 23 termes selon la volatilité de la série en question. La structure d'autocorrélation de la série observée est déterminée par l'autocorrélation des séries Y_t et e_t , structure qui, à son tour, influence l'estimation des composantes saisonnières, de tendance et intégrales. On peut estimer la structure de covariance de la série d'erreurs d'échantillonnage, e_t , d'après les données d'enquête au niveau des unités d'échantillonnage. Le calcul de ces estimations permet d'estimer la variance d'échantillonnage des séries estimatives de la tendance, de la saisonnalité et des irréguliers. À cet égard, diverses méthodes ont été proposées, par exemple, Steel et DeMel (1988) considèrent l'effet des filtres linéaires sur le spectre de la série d'erreurs d'échantillonnage, tandis que Wolter et Monsour (1981) adoptent une méthode fondée sur l'effet de filtres linéaires sur la fonction d'autocovariance. Sutcliffe (1993) adopte une méthode semblable axée sur une approximation linéaire de la méthode de X11. Pfefferman (1994) propose de produire une estimation de l'erreur d'échantillonnage directement d'après les séries chronologiques estimées. Ces méthodes ne modélisent pas

L'effet de divers plans de renouvellement sur la variance d'échantillonnage des estimations de la tendance

C.H. McLAREN et D.G. STEEL¹

RÉSUMÉ

Dans les domaines social et économique, de nombreuses séries chronologiques sont fondées sur des enquêtes par sondage à plan d'échantillonnage complexe. Or, le plan d'échantillonnage influence les propriétés de la série chronologique. Plus précisément, le chevauchement de l'échantillon d'une période à l'autre influence la variabilité de la série chronologique d'estimations basées sur des données d'enquête, ainsi que les estimations désaisonnalisées et les estimations de la tendance produites d'après ces séries chronologiques. La variante X11 de la méthode du US Census Bureau et le programme X11ARIMA, que l'on utilise couramment pour produire des estimations désaisonnalisées, permettent aussi de produire des estimations de la tendance. Le présent article décrit les effets de divers plans de chevauchement des échantillons sur la variance d'échantillonnage des estimations désaisonnalisées et des estimations de la tendance calculées d'après des séries chronologiques fondées sur des enquêtes par sondage.

MOTS CLÉS: X11; X11ARIMA; désaisonnalisation; estimation de la tendance; plan de renouvellement.

1. INTRODUCTION

Nombre de séries chronologiques importantes sont produites d'après les données d'enquêtes par sondage répétées dont le plan de chevauchement des échantillons d'une période à l'autre est complexe. Compte tenu de l'échantillonnage, la variabilité des séries chronologiques estimes tient, en partie, aux erreurs d'échantillonnage. Or, pour nombre de séries, ces erreurs sont une source importante de variabilité. Donc, le plan d'échantillonnage, en particulier le plan de la série chronologique d'estimations d'enquête.

L'étude des séries chronologiques se concentre de plus en plus sur l'évaluation des courbes sous-jacentes de variation ou des tendances par analyse des séries désaisonnalisées. La plupart des bureaux gouvernementaux de la statistique produisent des séries de données désaisonnalisées. Selon Kenny et Durbin (1982), les analystes de politiques déclarent souvent qu'ils s'intéressent davantage aux tendances sous-jacentes qu'aux fluctuations irrégulières des valeurs mensuelles non désaisonnalisées. Smith (1997) est du même avis. Pendant plus de dix ans, l'Australian Bureau of Statistics (ABS) a publié des séries d'estimations de la tendance obtenues par application de la méthode des moyennes mobiles de Henderson (MMH) aux séries désaisonnalisées pour en lisser les composantes irrégulières (ABS 1987). D'autres bureaux gouvernementaux de la statistique produisent des estimations de la tendance par diverses méthodes (Knowles 1997). Puisqu'on les produit par application de certains traitements à la série originale, les estimations désaisonnalisées et les estimations de la tendance sont également

influencées par les erreurs d'échantillonnage. Suivant Bell et Kramer (1999), la composante dominante de la variance des estimations désaisonnalisées est souvent celle due à l'erreur d'échantillonnage. Dans certains cas, les séries sont fondées sur des échantillons indépendants au fil du temps, mais, en général, les échantillons se chevauchent dans une certaine mesure d'une période à l'autre en vue de réduire les coûts et l'écart-type des estimations de la variation entre deux périodes consécutives (Kish 1998).

Un aspect essentiel de l'élaboration du plan d'échantillonnage d'une enquête répétée est le plan de renouvellement, autrement dit, le plan d'intégration d'une unité partiellement, au moins, au fil du temps, qui déterminera le chevauchement des échantillons. L'objectif de la présente étude est de déterminer les effets du plan de renouvellement choisi sur la variance d'échantillonnage des estimations désaisonnalisées et des estimations de la tendance obtenues par la variante X11 de la méthode du US Census Bureau (méthode Census X11) mise au point par Shiskin, Young et Musgrave (1967) et par la méthode X11ARIMA mise au point par Dagum (1980 et 1988). On se concentre ici sur le niveau et sur la variation d'une période à l'autre des estimations désaisonnalisées et des estimations de la tendance.

2. PLANS DE RENOUVELLEMENT

Considérons une série chronologique univariée dont les valeurs y_t , $t = 1, \dots, T$ proviennent d'une enquête par sondage répétée. La valeur observée au temps t est reliée à la valeur réelle de la série dans la population finie, Y_t , par

$$y_t = Y_t + e_t$$

¹ C.H. McLaren et D.G. Steel, School of Mathematics and Applied Statistics, University of Wollongong, NSW 2522, Australia. Courriel électronique: craigm@uow.edu.au, dsteel@uow.edu.au.

6. CONCLUSION

Nous avons examiné l'effet du recours à des variations de deux méthodes d'estimation du prix de vente médian des maisons vendues: l'estimation directe et l'interpolation linéaire. L'interpolation linéaire suppose un classement de données continues en casiers de largeur standard. Cette largeur peut être arbitraire, peut différer considérablement selon le domaine et peut changer à mesure que la distribution de l'échantillon évolue au fil du temps. La transformation linéaire fondée sur le troisième quartile a semblé corriger cette difficulté. En présence des données transformées, la largeur et l'emplacement des casiers dans la distribution changent en fonction des données.

Nos résultats empiriques indiquent que le choix de la méthode exerce une influence marquée sur des estimations de la variance pour la répétition de type MHS. Notre étude de simulation a permis d'examiner les propriétés des différentes procédures d'estimation de la médiane pour les estimations de la variance avec répétition de type MHS. Dans les quatre populations simulées, les médianes interpolées selon les données transformées (largeurs de casier qui dépendent des données) ont fourni les meilleurs résultats, habituellement dans une large mesure. Ce qui plus est, cette méthode réduit grandement la surestimation de la variance. L'utilisation de casiers de largeur 25 pour l'échelle transformée (41 casiers au total) a donné les meilleures estimations du prix de vente médian et de la variance, pour une répétition de type MHS, et c'est elle que nous recommandons pour la Survey of Construction.

La méthode recommandée offre plusieurs avantages. Tout d'abord, elle est adaptable. Elle fonctionne bien pour un choix de distributions, puisque la largeur des casiers eux-mêmes dépend de la distribution en question. Deuxièmement, elle permet d'économiser les ressources informatiques puisque l'on évite le tri de demi-échantillons. Troisièmement, les intervalles qui dépendent des données se laissent aisément intégrer à un logiciel généralisé de traitement des enquêtes. Enfin, elle donne de meilleures estimations et de meilleures estimations de la variance avec répétition de type MHS (du moins pour le prix de vente des maisons vendues). Nous estimons que ces résultats se laissent généraliser pour d'autres distributions continues également, mais il convient bien sûr de vérifier cette conclusion avec d'autres ensembles de données. On pourrait poursuivre la recherche, par exemple en examinant la relation entre la taille de l'échantillon et la précision des estimations de la médiane, en examinant d'autres tailles de casier et en observant la robustesse de la procédure recommandée à l'aide de diverses procédures d'estimation de la variance avec répétition.

REMERCIEMENTS

Les auteurs tiennent à remercier Elizabeth Huang et James Fagan du Bureau of the Census des États-Unis, de

BIBLIOGRAPHIE

- et de stimuler la discussion.
- même que deux examinateurs anonymes et le rédacteur associé de leurs commentaires anonymes et de versions antérieures du présent manuscrit, ainsi que J.N.K. Rao de l'Institut de statistiques de l'Université de Chicago. Il a subi un examen moins minutieux que les publications officielles du Bureau of the Census. La diffusion du présent rapport doit permettre d'informer les chercheurs intéressés par la question.
- DEGROOT, M. (1986). *Probability and Statistics*. Reading, MA: Addison-Wesley Publishing, Inc.
- FAY, R.E. (1989). Theory and application of replicate weighting for variance calculations. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.
- FAY, R.E. (1995). VPLX: Variance Estimation for Complex Surveys. Program Documentation. Non publié Bureau of the Census Report.
- HOAGLIN, D.C., et IGLEWICZ, B. (1987). Fine-tuning some resistant rules for outlier labeling. *Journal of the American Statistical Association*, 83, 1147-1149.
- JUDKINS, D.R. (1990). Fay's method for variance estimation. *Journal of Official Statistics*, 6, 223-239.
- KOVAR, J.G., RAO, J.N.K., et WU, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *La revue canadienne de statistique*, 16, 25-45.
- KOVACEVIC, M., et YUNG, W. (1997). Estimation de la variance des mesures de l'infériorité et de la polarisation du revenu - Étude empirique. *Techniques d'enquête*, 23, 47-59.
- LUERY, D.M. (1990). *Survey of Construction Technical Paper*. Ébauche non publiée Bureau of the Census documentation interne.
- NAYLOR, T.H., BALINTFY, J.L., BURDICK, D.S., et CHU, K. (1968). *Computer Simulation Techniques*. New York: John Wiley and Sons, Inc.
- RAO, J.N.K., WU, C.F.J., et YUE, K. (1992). Quelques travaux récents sur les méthodes de rééchantillonnage applicables aux enquêtes complexes. *Techniques d'enquête*, 18, 225-234.
- RAO, J.N.K., et SHAO, J. (1996). On balanced half-sample variance estimation in stratified random sampling. *Journal of the American Statistical Association*, 91, 343-348.
- THOMPSON, K.J. (1998). *Evaluation of Modified Half-Sample Replication for Estimating Variances for the Survey of Construction (SOC)*. Washington, DC: U.S. Bureau of the Census. (Rapport technique #ESM-9801, disponible Economic Statistical Methods and Programming Division).
- WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag, Inc.
- WOODRUFF, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 635-646.

Tableau 4
Le biais relatif et la stabilité relative pour des estimations de la variance et les taux d'erreur pour des intervalles de confiance de 90%

Population	Procédure d'estimation de la médiane	Plan à un degré sans grappes			Plan à deux degrés avec grappes		
		Biais relatif	Stabilité relative	Taux d'erreur	Biais relatif	Stabilité relative	Taux d'erreur
Financement conventionnel	SM	0,19	0,69	11,0%	0,11	0,58	15,1%*
	IO2000	0,25	0,35	6,9%*	0,25	0,37	9,0%
	IT4	0,21	0,32	7,0%*	0,19	0,33	9,3%
	IT25	0,06	0,25	10,0%	0,06	0,27	11,3%
Région 2	IT50	0,05	0,26	9,5%	0,05	0,28	12,1%*
	SM	0,57	1,24	7,3%*	0,41	1,07	7,9%*
	IO2000	0,33	0,44	6,9%*	0,23	0,35	8,6%
	IT4	0,30	0,42	7,0%*	0,17	0,30	8,7%
Midwest	IT25	0,15	0,41	10,1%	0,14	0,41	11,5%*
	IT50	0,16	0,40	9,8%	0,11	0,37	10,4%
	SM	0,15	0,42	9,0%	0,11	0,40	10,4%
	IO2000	0,31	0,42	6,7%*	0,28	0,39	7,5%*
Région 3	IT4	0,29	0,40	6,7%*	0,27	0,38	7,3%*
	IT25	0,02	0,28	11,0%	-0,01	0,27	10,8%
	IT50	0,01	0,29	11,1%	-0,02	0,28	11,9%*
	SM	0,39	0,98	8,9%	0,25	0,79	8,6%
Région 4	IO2000	0,32	0,42	6,2%*	0,31	0,41	5,2%*
	IT4	0,11	0,32	8,6%	0,10	0,31	7,6%*
	IT25	0,10	0,31	9,4%	0,09	0,30	7,5%*
	IT50	0,08	0,31	9,5%	0,08	0,31	8,3%*

4, les taux de couverture des trois procédures sont inférieurs à la valeur nominale. Toutefois, pour les casiers de largeur 4 et 25, un seul taux d'erreur est significativement supérieur à 10%; pour les casiers de largeur 50, deux de ces trois taux d'erreur sont significativement supérieurs à 10%. Toutes les médianes des données interpolées comportent des taux d'erreur significativement plus petits que la valeur nominale dans la population de la région 4; en conformité avec les résultats de l'autre population, les taux d'erreur pour les médianes interpolées selon les données originales sont les plus éloignés de 10%.

Dans les deux études, les médianes interpolées selon les données transformées offrent très nettement de meilleures propriétés d'estimation de la variance pour ce qui est du biais relatif et de la stabilité relative, peu importe la largeur du casier. De même, dans les deux études, les médianes interpolées selon les données transformées à l'aide de casiers de largeur 4 ou de largeur 25 offrent une excellente couverture des intervalles de confiance. Puisque les médianes interpolées selon les données transformées à l'aide de casiers de largeur 50 ou de largeur 25 ont donné les «meilleurs» estimateurs pour ce qui est de la racine de l'EQM et de l'EAM dans les deux études, il semble que l'utilisation de l'interpolation linéaire pour les données transformées comportant des casiers de largeur 25 soit la meilleure procédure d'estimation de la médiane pour ce qui est des propriétés d'estimation et d'estimation de la variance.

Les résultats pour la couverture des intervalles de confiance de chaque procédure d'estimation de la médiane ne sont pas aussi uniformes; ils varient selon le plan de sondage. En présence du plan à un degré sans grappes, les intervalles de confiance établis à partir des BT et des médianes interpolées selon les données originales comportent des valeurs extrêmement prudentes. Ici, le biais positif des estimations de la variance donne une largeur inutilement grande à ces intervalles, réduisant ainsi la possibilité de résultats intéressants. La couverture avec la médiane de l'échantillon est irrégulière. Certains de ces profils de couverture sont répétés dans le plan à deux degrés avec grappes. Encore une fois, la couverture avec la médiane de l'échantillon est irrégulière, et les taux de couverture pour les intervalles de confiance établis à partir des médianes interpolées selon les données originales sont supérieurs à la valeur nominale (bien significativement seulement). Le profil du taux d'erreur est bien différent pour les médianes interpolées selon les données transformées. Dans toutes les populations sauf celle de la région

Tableau 3
La racine de l'EQM empirique, l'erreur type (ET), le biais et l'erreur absolue moyenne (EAM)
pour les procédures d'estimation de la médiane

Population	Procédure d'estimation de la médiane	Echantillon à un degré non en grappes			Echantillon à deux degrés en grappes		
		SM	IO2000	IO1000	IT4	IT25	IT50
Financement conventionnel	SM	3 345	3 316	3 368	3 340	3 293	3 265
	IO2000	-12	161	-354	273	276	2606
	IO1000	2 671	2 698	2 642	3 431	3 378	3 305
	IT4	3 389	3 346	3 431	3 378	3 321	3 283
	IT25	3 374	3 341	3 420	3 364	3 311	3 275
Région 2 Midwest	SM	3 374	3 341	3 420	3 364	3 311	3 275
	IO2000	3 374	3 341	3 420	3 364	3 311	3 275
	IO1000	3 374	3 341	3 420	3 364	3 311	3 275
	IT4	3 374	3 341	3 420	3 364	3 311	3 275
	IT25	3 374	3 341	3 420	3 364	3 311	3 275
Région 3 Sud	SM	3 374	3 341	3 420	3 364	3 311	3 275
	IO2000	3 374	3 341	3 420	3 364	3 311	3 275
	IO1000	3 374	3 341	3 420	3 364	3 311	3 275
	IT4	3 374	3 341	3 420	3 364	3 311	3 275
	IT25	3 374	3 341	3 420	3 364	3 311	3 275
Région 4 Ouest	SM	3 374	3 341	3 420	3 364	3 311	3 275
	IO2000	3 374	3 341	3 420	3 364	3 311	3 275
	IO1000	3 374	3 341	3 420	3 364	3 311	3 275
	IT4	3 374	3 341	3 420	3 364	3 311	3 275
	IT25	3 374	3 341	3 420	3 364	3 311	3 275
IT50	SM	3 374	3 341	3 420	3 364	3 311	3 275
	IO2000	3 374	3 341	3 420	3 364	3 311	3 275
	IO1000	3 374	3 341	3 420	3 364	3 311	3 275
	IT4	3 374	3 341	3 420	3 364	3 311	3 275
	IT25	3 374	3 341	3 420	3 364	3 311	3 275

Lorsque nous avons examiné les propriétés d'estimation de la variance pour chaque procédure, les résultats ont été bien différents. Comme pour notre analyse des données empiriques, nous avions trois séries de résultats très distinctes. Le tableau 4 résume les trois mesures de comparaison pour les estimations de la variance des quatre populations. Les numérateurs pour le biais relatif et la stabilité et les taux de couverture se fondent sur 1 000 échantillons. Le dénominateur pour le biais relatif et la stabilité («vérité») se fondent sur 5 000 échantillons. L'astérisque (*) de la dernière colonne du tableau 4 indique que le taux d'erreur est appréciablement différent du taux d'erreur nominal de 0,10 en présence de l'approximation normale de la distribution binomiale pour le niveau de confiance de 90%.

Dans les deux études, les estimations de la variance des médianes interpolées selon les données transformées donnent les meilleurs résultats pour ce qui est du biais relatif et de la stabilité. En particulier:

- Les estimations de la variance des médianes interpolées selon les données transformées (IT4, IT25, IT50) comportent le biais relatif le plus petit. La différence liée à la méthode d'estimation est bien marquée dans

trois des quatre populations, où le biais relatif le plus important des médianes interpolées selon les données transformées représente moins de la moitié de la taille du plus petit biais relatif des médianes de l'échantillon. Ces résultats sont remarquablement solides pour le plan à deux degrés en grappes, puisque les estimations de la variance sont supposées comporter un biais vers le haut (voir la section 5.A).

Les estimations de la variance des médianes interpolées selon les données transformées ont donné, de façon générale, de meilleurs résultats que les médianes interpolées selon les données originales, bien que la différence ne soit pas aussi marquée que dans le cas du biais relatif. En général, la stabilité est proche avec les trois largeurs de casier pour les médianes interpolées selon les données transformées.

les médianes interpolées selon les données transformées linéairement, casiers de taille 50 (largeur de casier qui dépend des données)

Nous avons calculé $M(5)$, l'erreur quadratique moyenne empirique de la procédure d'estimation de la médiane à sous la forme

$$(5.1) \quad \hat{\theta}_2(\zeta_j) + \hat{\text{bias}}_2(\zeta_j) =$$

Nous avons calculé l'erreur absolue moyenne (MAE) de chaque procédure d'estimation de la médiane à sous la forme

suivant la définition de DeGroot (1986, 209-211). Afin de comparer les propriétés d'estimation

2.

d'une procédure à l'autre pour chaque population. Comparaison des propriétés d'estimation de la variance avec répétition de type MHS des procédures d'estimation de la médiane.

B. Résultats

θ_{11} est l'extrême inférieure d'un intervalle de confiance de 90% et

Taux d'erreur Nombre d'échantillons où $(\zeta^p > \theta_{Li}$ ou ζ^p

La médiane, le troisième quartile et la largeur des casiers pour l'échelle originale des données simulées transformées

Population	Médiane	Q_3	4	25	Largeur de casier
Financement conventionnel	167 173	222 263	889	5 557	11 113
Midwest (région 2)	151 312	210 647	843	5 266	10 532
Sud (région 3)	133 745	180 868	723	4 522	9 043
Ouest (région 4)	162 130	214 320	857	5 358	10 716

taille, cette étude ne tient pas compte de l'échantillon-nage PPT de la SOC et ne comprend pas le groupement des unités du premier degré. La répétition de type MHS fait appel aux unités d'échantillonnage (UPB) du premier degré à l'intérieur de la même strate. Les poids répétés ne rendent pas compte d'importantes fractions d'échantillonnage au premier degré de la sélection comme le recommande Wolter (1985, 122), de sorte que les estimations de la variance comportent probable-ment toutes un biais vers le haut.

Nous n'avons pas tenté de simuler l'échantillon SUP des permis pour quatre unités de logement ou moins dans des UPB non autoreprésentatives et des bureaux de permis non autoreprésentatifs (échantillon à trois degrés, 25% à peu près de l'échantillon de la SOC); l'échantillon SUP des permis pour cinq unités de logement ou plus (2% environ de l'échantillon de la SOC); ou l'échantillon NP des DD (5% environ de l'échantillon de la SOC). L'échantillon à trois degrés, bien que non négligeable dans la SOC, est rarement utilisé dans d'autres enquêtes du Bureau of the Census, et les deux autres secteurs du plan de la SOC ne comptent pas suffisamment dans les estimations pour justifier une étude distincte.

Afin d'examiner la précision de chaque procédure d'estimation de la médiane pour des échantillons répétés, nous avons estimé les erreurs quadratiques moyennes (EQM) et les erreurs absolues moyennes (EAM) empiriques des 5 000 échantillons en fonction de ce qui suit:

- ME:** la médiane de l'échantillon pour chaque demi-échantillon
- IO2000:** les médianes interpolées selon les données originales, casiers de taille 2 000 (largeur de casier fixe)
- IO1000:** les médianes interpolées selon les données originales, casiers de taille 1 000 (largeur de casier fixe)
- IT4:** les médianes interpolées selon les données transformées linéairement, casiers de taille 4 (largeur de casier qui dépend des données)

Tableau 1
Caractéristiques des populations simulées et taille des échantillons stratifiés

Population	Distribution	θ	σ	ζ	Corrélation (stratificateur, prix de vente)	Taille de la population	Taille de l'échantillon	Paramètres de prix de vente	
								p	N
Financement conventionnel	normale logarithmique	27 578	0,4895	11,84	0,57030	25 150	500	Midwest	normale logarithmique
		31 801	0,5957	11,69	0,55835	6 500	150	Sud	normale logarithmique
		29 414	0,5549	11,55	0,55929	14 550	300	Ouest	normale logarithmique
		53 781	0,5822	11,59	0,55525	11 550	250		

standard servant à créer la variable du prix de vente. Les percentiles, l'asymétrie de l'échantillon et l'aplatissement de l'échantillon de la variable du prix de vente de chaque population simulée se rapprochaient beaucoup de la statistique correspondante de la population originale, surtout lorsque les valeurs aberrantes étaient supprimées à l'aide de la règle robuste des clôtures extrêmes décrite dans Hoaglin et Iglewicz (1987). La taille de chaque population était le N estimé à partir des populations de l'échantillon. Les paramètres de modèle, les corrélations de l'échantillon (entre les prix de vente simulés et la variable stratifiante), la taille de la population (N) et la taille des échantillons (n) sont présentées dans le tableau 1.

Après avoir créé les populations finies, nous avons formé 50 strates de taille égale dans chaque population, puis nous avons tiré deux séries d'échantillons pour deux plans d'enquête différents:

- Le premier s'inspire de l'échantillon SUP des permis pour quatre unités de logement ou moins des bureaux de permis dans des UPB autoreprésentatives (28% à peu près de l'échantillon de la SOC). Dans cette étude, nous avons tiré 5 000 échantillons stratifiés aléatoires sans remise pour chaque population simulée à l'aide du même taux d'échantillonnage dans chaque strate. Afin d'effectuer une répétition de type MHS, nous avons tiré l'échantillon à l'intérieur de chaque strate selon la variable stratifiante et nous avons divisé systématiquement l'échantillon en deux panels.
- Le deuxième plan s'inspire de l'échantillon SUP des bureaux de permis non autoreprésentatifs dans des UPB autoreprésentatives et des bureaux de permis pour quatre unités de logement ou moins des bureaux de permis non autoreprésentatifs dans des UPB autoreprésentatives (40% à peu près de l'échantillon de la SOC). Dans cette étude, nous avons tiré 5 000 échantillons à deux degrés pour chaque population simulée. Le premier degré est un échantillon stratifié aléatoire sans remise de deux UPB par strate ($N_h = 5$). Le deuxième degré est un échantillon systématique d'unités à l'intérieur des UPB. Puisque toutes les UPB ont la même

L'intervalle ($\bar{p} \pm \text{ET}(\bar{p})$) sur la distribution cumulative des fréquences de façon à fournir la limite inférieure d'un intervalle de confiance de 52,86% pour la médiane ($\text{L'ET}(\bar{p})$ peut être estimée à l'aide de méthodes de répétition). L'ET de la médiane est alors estimée par soustraction. Cette méthodologie a été plus ou moins fructueuse dans le passé d'après les analystes de l'enquête SOC.

4. RÉSULTATS DES DONNÉES EMPIRIQUES

Initialement, nous avons utilisé quatre mois des données de l'échantillon de la SOC afin d'examiner la variance des méthodes d'estimation de la médiane pour le prix de vente des maisons vendues: mars 1997, mai 1997, juin 1997 et juillet 1997. Nous avons préparé des médianes selon la région et selon le type de financement. Nous avons utilisé les mêmes poids que pour les systèmes de variance et d'estimation de la production de la SOC (stratifiés à posteriori pour l'échantillon SUP et non biaisés pour l'échantillon NP), les données des deux enquêtes étant groupées en vue de l'obtention de médianes. Chaque série d'estimations de la variance a été produite à l'aide de 200 répétitions.

Nous avons observé que les six méthodes d'estimation de la médiane ont produit des estimations très semblables, dominant lieu toutefois à trois séries distinctes d'ET: une série pour la médiane de l'échantillon, une série pour les médianes interpolées selon les données originales (largeur de casier fixe) et une série pour les médianes interpolées selon les données transformées (largeur de casier qui dépend des données). Il n'y avait pas de lien net entre la largeur du casier et les estimations de l'ET pour les deux séries de médianes interpolées. En effet, pour un même type de données (originales ou transformées), les ET étaient toutes très rapprochées. De toute évidence, il y avait une transformation linéaire et un effet d'interpolation. Aucune des méthodes d'estimation de la médiane n'a donné des erreurs types ressemblant aux erreurs types publiées, de sorte que l'on ne pouvait pas conclure qu'il y avait uniformité des données publiées.

De plus, il semblerait que les ET publiées de la méthode de Woodruff sont sous-estimées ou du moins inappropriées pour le plan d'échantillonnage utilisé. Kovar, Rao et Wu (1988) ont comparé les ET de Woodruff et celles de la méthode BRR, et ils ont trouvé que les deux méthodes avaient des propriétés semblables, sauf dans le cas des échantillons stratifiés, où les strates se fondent sur des variables séparées fortement corrélées (comme le plan de la SOC). Dans ce cas, l'ET de Woodruff est souvent trop petite, et ils ont conclu que les méthodes BRR (sic) sont plus robustes vis-à-vis de diverses structures démographiques, puis que l'erreur est extraite directement des répétitions. Lorsque les ET de Woodruff du système de production faisaient appel à l'ET(\bar{p}) calculée directement, les ET de Woodruff étaient généralement plus petites que les ET répétées.

Les résultats empiriques nous ont laissés perplexes. Nous avons trois séries distinctes d'estimations de la variance et aucun «étalon-or» permettant de les mesurer. Puisque nos résultats empiriques n'étaient pas convaincants, nous avons mené une étude de simulation de Monte Carlo afin d'évaluer les propriétés des estimations de la variance de type MHS préparées à l'aide des différents estimateurs de la médiane.

5. COMPARAISON DES ÉTUDES DE SIMULATION

A. Procédure pour les études de simulation

Nous avons créé quatre populations artificielles finies fondées sur une analyse des données de quatre populations de l'échantillon de la SOC: une population du type de financement (financement conventionnel) et trois populations régionales (Midwest = région 2, Sud = région 3 et Ouest = région 4). Ces populations représentaient divers types des populations de la SOC servant à produire des estimations. À noter que la population du type de financement de la SOC n'est pas indépendante des populations des régions de la SOC.

Afin d'obtenir une approximation de la population finie du prix de vente des maisons vendues, nous avons créé w_i enregistres pour chaque unité d'échantillonnage i , où w_i est le poids d'échantillon associé à l'unité i . Des distributions normales logarithmiques ont permis d'obtenir une approximation de la distribution du prix de vente des maisons vendues à une seule unité. La distribution normale logarithmique comporte la fonction de densité

$$f(y) = \frac{1}{1} \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2} \left(\frac{\log(y) - (\theta - \zeta)}{\sigma}\right)^2\right) \text{ pour } \theta < y < \infty$$

où θ est le paramètre de seuil, ζ est le paramètre d'échelle et σ est le paramètre de forme.

À l'aide de nos modèles, nous avons créé quatre populations simulées finies à deux variables ayant une corrélation prévue de $\rho = 0,6$ à l'aide de la méthode décrite dans Naylor, Baintiff, Burdick et Chu (1968, 99). La première des deux variables de chaque population représentait le prix de vente des maisons vendues, et elle a été obtenue par la création d'une variable normale aléatoire avec moyenne ζ et variance σ^2 à l'aide des paramètres déterminés ci-dessus, suivie d'une élévation à la puissance et d'un décalage à l'aide de paramètres d'emplacement appropriés (θ). La deuxième variable a servi à former des strates et des grappes de premier degré. Cette variable comportait une distribution normale standard marginale; elle a été obtenue par la création indépendante d'une deuxième valeur normale aléatoire standard, multipliée par 0,8, ce terme étant additionné à $0,6 \times$ la variable aléatoire normale

concevable que, pour des catégories de financement spécial ou certaines régions, le prix de vente médian des maisons vendues se rapproche de 550 000\$, ce qui entraînerait l'échec de l'interpolation.

Afin d'évaluer la deuxième option d'interpolation (médiannes interpolées selon des données transformées), nous avons eu recours à trois ensembles de largeurs de casier: casiers de taille 4, 25 et 50. Les casiers de taille 4 ont été choisis semblables aux casiers de taille 2 000 pour ce qui est du nombre de casiers. Il y a 250 casiers de taille 4 pour les données transformées inférieures à Q_3 , et un casier plus grand contenant toutes les données supérieures à Q_3 . La sélection des largeurs 25 et 50 a été relativement arbitraire: nous avons choisi le casier de taille 50 afin d'avoir un total de 20 casiers pour les données inférieures à Q_3 ; nous avons choisi le casier de taille 25 pour examiner l'effet du doublement du nombre de casiers et de la réduction de moitié de la largeur des casiers pour les données inférieures à Q_3 . La médiane des données transformées est toujours inférieure à 1 000, de sorte que le dernier classement de données transformées est toujours de (1 000 à maximum). Ainsi, par définition, le dernier casier contient jusqu'à 25% des données, et il est considéré comme plus large que les autres casiers.

B. Estimation de la variance

Nous avons utilisé la méthode de répétition MHS (modifiée avec demi-échantillon) (Fay 1989 et Judkins 1990) pour estimer la variance d'une médiane tel que mentionné dans la documentation publiée (par exemple, Rao, Wu et Yue (1992); Rao et Shao (1996); Kovacevic et Yung (1997) pour la répétition équilibrée; Judkins (1990) pour la répétition de type MHS). La répétition de type MHS est une variation de l'estimation équilibrée «traditionnelle» de la variance avec demi-échantillon décrite par Wolter (1985, 110-152). La répétition BRR (équilibrée avec demi-échantillon) est une méthode d'estimation de la variance conçue pour un plan de sondage de deux UPB par strate. Selon la méthode BRR, une répétition avec demi-échantillon est effectuée par sélection d'une unité de chaque paire et par pondération de l'unité sélectionnée par 2 (de façon qu'elle représente les deux unités). Ainsi, des estimations pour chaque UPB sont comprises dans chaque répétition même si la moitié d'entre elles ont une pondération zéro. Les répétitions (demi-échantillons) sont précisées à l'aide d'une matrice de Hadamard. On trouvera dans Wolter (1985, 114-115) une description détaillée de la procédure de répétition axée sur des matrices de Hadamard. La répétition de type MHS fait appel à des poids de répétition de 1,5 et de 0,5 au lieu de 2 et 0. L'erreur type d'une estimation de la médiane axée sur une répétition de type MHS est donnée par

$$SE(\text{Med}) = \sqrt{\frac{1}{4} \sum_{r=1}^R (\text{Med}_r - \text{Med}_0)^2} \quad (2.3)$$

où l'indice r désigne l'estimation de la médiane r répétée ($r=1,2,...,R$) et l'indice 0 désigne l'estimation de la

Ni le plan SUP ni le plan NP n'est un plan de sondage à l'EQM des estimations répétées est trop faible d'un facteur de 1/(1-0,5)². Voir Judkins (1990).

deux unités d'échantillonnage par strate. Au premier degré, on sélectionne une UPB par strate. Les deuxième et troisième degrés sont des échantillons systématiques, et souvent une seule unité par strate a été sélectionnée au deuxième degré. Une façon courante d'aborder le problème suscitée par une unité d'échantillonnage par strate consiste à «diviser» les unités d'échantillonnage autoreprésentatives en deux panels par unité d'échantillonnage à l'aide de la méthode d'échantillonnage originale; former des strates groupées par appariement de deux (ou trois) unités d'échantillonnage non autoreprésentatives «semblables»; et

appliquer la stratégie du demi-échantillon de façon que les éléments des demi-échantillons soient des panels à l'intérieur d'unités d'échantillonnage pour les unités d'échantillonnage autoreprésentatives, et qu'ils soient les unités d'échantillonnage de premier degré (UPB) à l'intérieur de strates groupées pour les unités d'échantillonnage non autoreprésentatives.

Le système actuel de variance pour la production de la SOC fait appel à un estimateur de Keyfitz (estimateur de différence apparié) pour un échantillon non autoreprésentatif et à un estimateur de formule d'échantillonnage apparié pour un échantillon autoreprésentatif afin de produire des estimations uniformes de la variance (Luey 1990). Puisque les méthodologistes de la SOC avaient déjà groupé des strates non autoreprésentatives pour leur estimation de différence apparié, une application de type BRR était un prolongement logique de la structure préexistante d'estimation de la variance. Pour une répétition de type MHS, nous tirons les permis à l'intérieur de groupes d'unités d'échantillonnage prédéterminés en unités autoreprésentatives selon le code géographique et la date d'autorisation, et nous divisons systématiquement l'échantillon ordonné en deux panels comme le suggère Wolter (1985, 131). Bien que ce soit essentiellement la seule stratégie disponible pour le plan de sondage de la SOC, cette méthode risque de ne pas fournir des estimations correctes de la variance puisque des unités des deux panels sont corrélées (dans la méthode originale avec demi-échantillon, les deux UPB de la strate sont supposées indépendantes). On trouvera des détails supplémentaires sur l'attribution des répétitions dans Thompson (1998).

Le système de production de la SOC fait appel à la méthode de Woodruff (Woodruff 1952) afin d'évaluer l'erreur type d'une médiane. La méthode de Woodruff fait appel à l'erreur type estimative d'une proportion p ($p = 0,50$ pour une estimation de la médiane) avec projection de

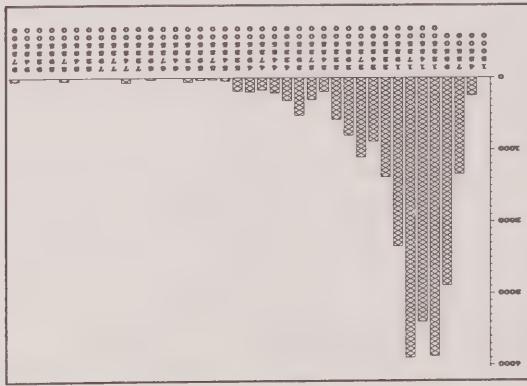


Figure 1: Distribution originale du prix de vente des maisons vendues ayant une largeur conventionnelle de casier de financement (largeur de casier = 25 000\$).

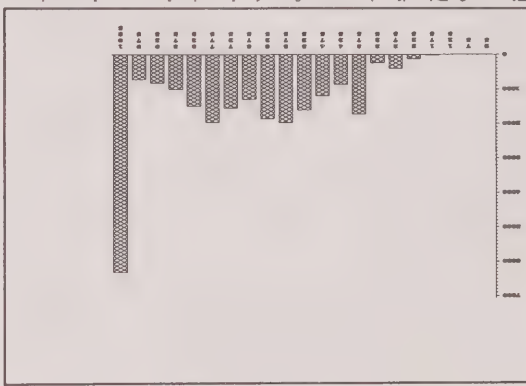


Figure 2: Distribution transformée du prix de vente des maisons vendues moyennant un financement conventionnel pour une largeur de casier correspondant à une largeur de casier = 50 de l'échelle non transformée = 1 250\$.

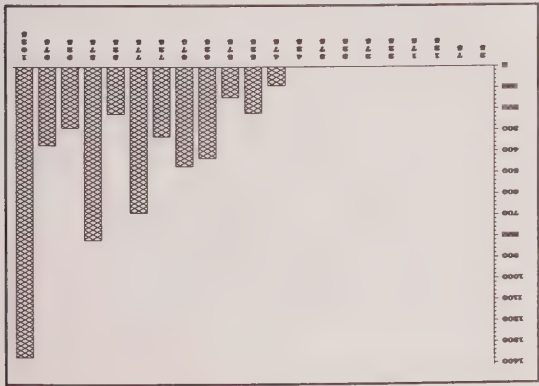


Figure 4: Distribution transformée du prix de vente des maisons vendues moyennant des prêts de la FHA (largeur de casier = 50). Largeur de bande dans l'échelle non transformée = 6 250\$.

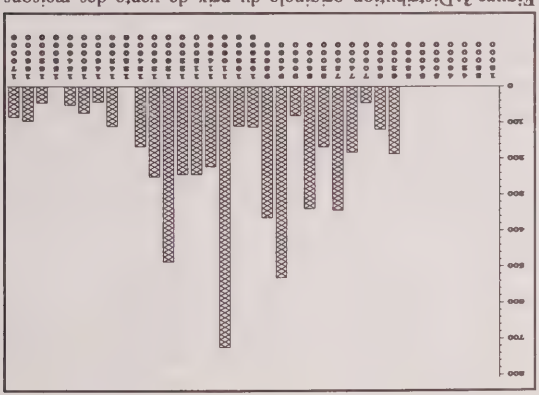


Figure 3: Distribution originale du prix de vente des maisons vendues moyennant des prêts de la FHA (largeur de casier = 4 000\$).

Afin d'évaluer la première option d'interpolation (médianes interpolées selon des données originales), nous avons utilisé deux ensembles de largeurs de casier (tailles de classement): casiers de taille 2 000\$ (la même largeur de casier utilisée pour le système actuel d'estimation de la variance pour la production) et casiers de taille 1 000\$. [Nota: le logiciel VPLX d'estimation de la variance ne permet aucune taille de casier inférieure à 1 000\$ puisque le nombre de catégories dépassait l'éventail admissible.] Après avoir examiné plusieurs mois d'estimations du prix de vente pour l'ensemble des États-Unis, nous avons supposé que le prix de vente médian serait toujours supérieur à 36 000\$ et inférieur à 550 000\$; le premier classement de données originales est donc toujours de (minimum à 35 999), le dernier classement de données originales étant toujours de (550 000 à maximum): cela donne 257 casiers de taille 2 000\$ ou 514 casiers de taille 1 000\$, plus un casier de taille 36 000\$ et un casier dont la largeur dépend de la plus grande observation de l'échantillon. Un problème évident lié à l'emploi de ces casiers est l'effet éventuel de l'inflation. Il est

La figure 3 présente un histogramme de la distribution originale des maisons vendues moyennant des prêts de la FHA, avec une largeur de casier de 4 000\$ (encore une fois, la largeur de casier est choisie pour des raisons de présence). La médiane de cette distribution est de 1 082 80\$, et on a 124 990 pour Q_3 . La figure 4 présente l'histogramme de la distribution transformée linéairement avec une largeur de casier de 50. Dans cet exemple, les casiers transformés de largeur 50 correspondent à des casiers de largeur 6 250\$ dans l'échelle originale, et les casiers des données transformées de largeur 4 auraient une largeur approximative de 500\$ dans l'échelle non transformée originale.

Les figures 1 à 4 montrent la souplesse des casiers élaborés pour des données transformées linéairement. La taille de casier dans l'échelle non transformée peut être plus grande ou plus petite suivant l'écart des données. De plus, la taille des échantillons de casier dépendant des données est moins variable comparativement à celle qui est associée à des casiers fixes.

N = le nombre total estimatif des éléments de la

population

cf = la fréquence cumulative pour tous les intervalles

précédant le casier contenant la médiane

f_j = la fréquence de la catégorie médiane (nombre

total estimatif des éléments de la population de

l'intervalle contenant la médiane)

i = la largeur du casier contenant la médiane.

C'est là la méthode utilisée pour le système actuel d'es-

timation de la variance pour la production de la SOC pour

les estimations mensuelles, et c'est également la méthode

d'interpolation linéaire utilisée pour le progiciel VPLX.

Nous avons considéré deux options pour le réglage de la

taille de catégorie (largeurs de casier) pour l'interpolation.

La première option consiste à élaborer des casiers en fonc-

tion de la caractéristique particulière à l'étude selon les don-

nées originales. La deuxième option comporte une trans-

formation linéaire des données selon une échelle standard,

un ensemble standard de casiers étant alors utilisé pour

chaque caractéristique. Nous avons fait appel à la transfor-

mation linéaire ci-dessous:

$$X'_i = X_i * \frac{\sigma_3}{1,000} \quad (3.2)$$

où σ_3 est le troisième quartile de la distribution de

l'échantillon (estimé à l'aide des observations ordonnées et

du poids d'échantillon décrits à la section 2.A.1). La

médiane interpolée du X' est multipliée par ($\sigma_3/1,000$) de

façon à donner une médiane estimative selon l'échelle

originale. [Si la distribution comporte des valeurs négatives

(par exemple, une distribution du revenu net), on utilise

alors $X'_i = (X_i - X^{(1)}) * 1,000/\sigma_3(X'_i - X^{(1)})$, où $X^{(1)}$ est la

statistique d'ordre un et $\sigma_3(X'_i - X^{(1)})$ est calculé à partir de

la distribution de $(X'_i - X^{(1)})$. Pour obtenir une médiane

estimative selon l'échelle originale, on multiplie la médiane

interpolée par $(\sigma_3(X'_i - X^{(1)})/1,000$ et on ajoute $X^{(1)}$.]

Cette procédure correspond à la simple division de l'échan-

tilon original de 0 à σ_3 en x casiers de largeur égale, le

reste des données étant placé dans un casier qui, de par sa

conception, est beaucoup plus grand que les autres.

Cette procédure est conçue en fonction de distributions

symétriques ou positivement asymétriques (comme c'est ha-

bituellement le cas des données économiques). Les données

qui se trouvent dans le dernier casier ne servent pas à esti-

mer la médiane parce qu'elles sont plus grandes que σ_3 ,

jugé éloigné de la médiane. Si l'on fondait la transformation

linéaire sur σ_1 (le premier quartile), le casier qui contient

la médiane pourrait être très rapproché du casier inférieur

de la distribution. Dans un tel cas, la différence de varia-

bilité entre une médiane interpolée et la médiane de

l'échantillon serait petite.

Le fait d'utiliser les données originales afin d'élaborer

les médianes permet de préparer des estimations et des ET

prêtes pour la production. Il est difficile, toutefois, de déter-

miner la largeur de casier fixe appropriée. Plus la largeur

des casiers est petite (largeur 1), et plus les estimations de

la variance sont instables. L'augmentation de la largeur des

casiers fait augmenter le biais de l'estimation attribuable à

l'interpolation. La taille «optimale» du casier établit un équi-

libre entre le biais et la stabilité de l'estimation de la va-

riance. Malheureusement, la largeur optimale du casier ne

reste pas nécessairement constante d'un échantillon à l'au-

tre. Souvent, la distribution évolue au fil du temps, et la lar-

geur/emplacement des casiers dans l'échantillon devrait re-

fléter ce changement d'échelle. De plus, la largeur optimale

du casier peut différer pour diverses valeurs d'une variable

de classement; par exemple, la largeur optimale du casier

pour le prix de vente dans le Midwest diffère probablement

de la largeur optimale du casier pour le prix de vente dans

le Sud.

La transformation linéaire a été motivée par l'intérêt

pour une largeur de casier dépendant de l'échantillon. Les

largeurs de casier «standard» utilisées pour les données

transformées inférieures à σ_3 ne sont pas standard pour

l'échelle non transformée; la largeur du casier dépend des

données. L'utilisation des données transformées linéaire-

ment suppose un travail comptable accru pour ce qui est des

constantes d'échelle, mais permet facilement de changer

l'échelle et la forme de la distribution.

Les figures 1 à 4 montrent l'effet de la transformation li-

néaire sur la largeur des casiers et leur emplacement pour

deux distributions. Les figures 1 et 2 illustrent une distribu-

tion comportant un large éventail de valeurs de données, y

compris quelques observations très grandes. Les figures 3

et 4 illustrent une distribution constituée surtout de petites

valeurs de données.

La figure 1 présente un histogramme de la distribution

originale de maisons vendues moyennant un financement

conventionnel, avec une largeur de casier de 25 000\$ [Nota:

la taille du casier a été choisie purement pour faciliter la

présentation, puisqu'il s'agit d'une distribution allongée.]

La médiane de cette distribution est de 167 130\$, avec un σ_3

de 225 000\$. La figure 2 présente l'histogramme de la dis-

tribution transformée linéairement avec une largeur de ca-

sier de 50. Dans cet exemple, les casiers transformés de lar-

geur 50 correspondent à des casiers de largeur 11 250\$ dans

l'échelle originale (225 000\$/1 000\$*50). On se souven-

dra que la taille des données originales à l'étude

est de 1 000\$ et de 2 000\$. Ainsi, les casiers de données

transformées de largeur 4 auraient une largeur de 900\$ dans

l'échelle originale non transformée. On remarquera l'impor-

tañte «pointe» du dernier casier, qui contient toutes les va-

leurs de l'échantillon supérieures à σ_3 .

Ces figures illustrent également les différences de distri-

bution des tailles d'échantillon d'un casier à l'autre pour les

deux méthodes. L'utilisation de la largeur de casier fixes

avec les données originales entraîne des tailles assez varia-

bles d'échantillon de casier (voir la Figure 1). Par contre, de

par leur conception, les tailles d'échantillon à l'intérieur des

casiers qui dépendent des données sont beaucoup plus uni-

formes pour tous les casiers sauf le dernier (voir la figure

2).

présentée à la section 5. Nos conclusions et nos recommandations se trouvent à la section 6.

2. PLAN DE SONDAGE DE LA SOC

L'univers de la SOC contient deux sous-populations: des zones locales qui exigent des permis de construction et des zones locales qui n'en exigent pas. Les unités d'échantillonnage de la SOC tirées de la première sous-population comprennent l'enquête initiale Survey of the Use of Permits (SUP), et celles tirées de la deuxième sous-population, l'enquête Nonpermit Survey (NP). L'échantillon de la SUP constitue la plus grande partie de l'estimation de la SOC. Les deux sont des échantillons probabilistes à plusieurs degrés stratifiés selon des variables ayant une corrélation prévue élevée avec les statistiques clés de l'enquête: les mises en chantier, les travaux achevés et les ventes.

Le premier degré du tirage des échantillons SUP et NP est un sous-échantillon des unités primaires d'échantillonnage (UPB) du plan de 1980 de la CPS (Current Population Survey), qui sont des parcelles de terrain adjacentes ayant des limites bien définies. Ainsi, les deux enquêtes comportent les mêmes UPB, mais pour le reste les échantillons sont indépendants. Les UPB ont été stratifiées à l'intérieur de la région selon la population pondérée de 1980 des 16 ans et plus, les permis résidentiels pondérés de 1982 et le pourcentage de logements dans des zones sans permis. Dans la mesure du possible, les strates sont constituées d'UPB du même État ayant le même caractère métropolitain. On a tiré une UPB par strate. Des UPB autoréprésentatives (AR) ont été incluses dans l'échantillon avec certitude (la strate est constituée d'une UPB). Des UPB non autoréprésentatives (NAR) ont été tirées moyennant une probabilité proportionnelle à la taille (PPT) à partir de strates comportant plus d'une UPB.

Le deuxième degré du tirage de l'échantillon SUP est un échantillon systématique stratifié d'endroits émettant des permis à l'intérieur des UPB de l'échantillon (tirage une fois par décennie). Ces endroits ont été stratifiés en fonction d'une moyenne pondérée du rapport entre l'émission de permis pour l'année i et le total des permis émis aux États-Unis pour l'année i ($\bar{q} = 78, 81, 82$). Dans de nombreux cas, on a tiré une seule unité pour le deuxième degré. Le troisième degré du tirage de l'échantillon SUP est exécuté mensuellement: chaque mois, des représentants locaux (RL) sélectionnent un échantillon systématique de permis de construction provenant des bureaux de permis de chaque endroit échantillonné qui émet des permis. On sélectionne systématiquement des permis de construction d'un échantillon global de type un pour quarante; des permis de construction de cinq unités ou plus sont inclus avec certitude. Les échantillons du troisième degré sont indépendants selon le mois; ceux des premier et deuxième degrés ne le sont pas.

3. MÉTHODOLOGIE

A. Procédures d'estimation de la médiane

1. Médiane de l'échantillon
 2. Interpolation linéaire
- Une façon d'estimer la médiane d'une population consiste à calculer la médiane de l'échantillon à partir de données non groupées, le poids de l'échantillon permettant de situer la médiane. Cette stratégie est recommandée dans Kovar, Rao et Wu (1988) et dans Rao et Shao (1996). Cette procédure se fonde sur les étapes ci-dessous:
- trier les observations de l'échantillon dans l'ordre ascendant;
 - accumuler la somme des poids d'enquête associés;
 - sélectionner la première observation pour laquelle la somme associée des poids dépasse la moitié du poids total.

Une autre façon d'estimer la médiane d'une population consiste à grouper les données d'échantillon et à y interpoler la médiane de l'échantillon. Woodruff (1952) fournit la formule ci-dessous pour l'interpolation linéaire de la médiane d'un échantillon:

$$\hat{M} = F^{-1}\left(\frac{1}{2}N\right) \approx l + \left(\frac{\frac{1}{2}N - cf}{f_i}\right) * (i) \quad (3.1)$$

où

- F = la fréquence cumulative de la caractéristique axée sur des poids d'échantillon
- l = la limite inférieure du casier contenant la médiane

L'estimation et l'estimation de la variance par répliques des prix de vente médians des maisons vendues

KATHERINE J. THOMPSON et RICHARD S. SIGMAN¹

RÉSUMÉ

Le Bureau of the Census des États-Unis publie des estimations de la médiane pour plusieurs caractéristiques des maisons neuves, une estimation clé étant le prix de vente des maisons vendues. Ces estimations sont calculées à partir de données acquises par voie d'interview des constructeurs de maisons dans le cadre de l'enquête SOC (Survey of Construction). La SOC est une enquête probabiliste à plusieurs degrés dont le plan d'échantillonnage se prête bien à la méthode MHS (répétition modifiée avec demi-échantillon) d'estimation de la variance. Des documents publiés appuient l'application de la méthode MHS à la médiane d'échantillons répétés en vue de l'estimation de la variance d'échantillonnage d'une médiane. Il existe cependant plusieurs avantages, du point de vue des calculs, à utiliser des données groupées en vue de l'estimation des médianes, une interpolation linéaire étant utilisée dans l'intervalle des données groupées comportant la médiane. À l'aide de données d'enquête et de populations finies simulées, nous avons comparé l'effet de l'absence de groupement (donc la médiane de l'échantillon), le groupement avec intervalles de taille fixe et le groupement avec intervalles dont la taille dépend des données sur les médianes et sur les estimations MHS connexes de la variance. Nous avons examiné les erreurs quadratiques moyennes et les erreurs absolues moyennes des estimations de la médiane de même que le biais relatif et la stabilité des estimations de la variance et la couverture des intervalles de confiance connexes. Nous avons observé que les intervalles dont la taille dépend des données donnent des estimations de la variance comportant le biais le moins élevé, la meilleure stabilité et les meilleurs intervalles de confiance.

MOTS CLÉS : Médiane; répétition modifiée avec demi-échantillon; Survey of Construction.

1. INTRODUCTION

Le Bureau of the Census des États-Unis publie des estimations de la médiane pour plusieurs caractéristiques des maisons neuves, une estimation clé étant le prix de vente des maisons vendues. Ces estimations sont calculées à partir de données acquises par voie d'interview des constructeurs de maisons dans le cadre de l'enquête SOC (Survey of Construction). La SOC est une enquête probabiliste à plusieurs degrés dont le plan d'échantillonnage se prête bien à la méthode MHS (répétition modifiée avec demi-échantillon) pour des raisons qui sont expliquées à la section 3.B. Le Bureau of the Census fera bientôt passer la SOC des systèmes actuels d'estimation et d'estimation de la variance à son système remanié, le Standardized Economic Processing System (SEPS). À ce moment-là, la procédure actuelle d'estimation non répétée de la variance pour la SOC sera remplacée par la méthode MHS d'estimation de la variance avec répétition (Thompson 1998). Puisque la méthodologie d'estimation de la variance pour la SOC évolue, nous avons décidé de réexaminer la méthodologie d'estimation de la médiane pour des données continues. Il s'agissait de trouver une méthode d'estimation de la médiane ayant de bonnes propriétés d'estimation et d'estimation de la variance, compte tenu de la répétition de type MHS.

Nous avons considéré deux méthodes d'estimation de la médiane. La première fait appel aux poids d'échantillon pour estimer la médiane à l'aide de fonctions empiriques de

distribution cumulative. La deuxième méthode fait appel à une interpolation linéaire de données continues groupées pour obtenir une approximation de la médiane. Cette dernière méthode est mise en œuvre en VPLX (Variances from Complex Survey, Fay 1995), qui est le progiciel d'estimation de la variance avec répétition élaboré au Bureau of the Census.

Le calcul direct des médianes d'échantillon peut exiger un travail de calcul intense parce qu'il exige des tris distincts pour chaque valeur d'une variable de classement donnée. Une autre méthode d'estimation consiste à grouper les données continues en intervalles discrets (appelés casiers) et à utiliser une interpolation linéaire pour l'intervalle comportant la médiane. Pourvu que la distribution des données soit à peu près uniforme pour l'intervalle comportant la médiane, l'interpolation fournit une bonne approximation et exige beaucoup moins de ressources informatiques. Toutefois, la largeur et l'emplacement optimaux des casiers peuvent différer selon le domaine et risquent d'évoluer dans le temps à mesure que la distribution des échantillons change.

Dans le présent exposé, nous comparons six méthodes d'estimation de la médiane, compte tenu de la méthode de répétition MHS: la médiane de l'échantillon et cinq variations axées sur une interpolation linéaire. On trouvera à la section 2 un bref aperçu du plan de sondage de la SOC. La méthodologie générale est expliquée à la section 3. La section 4 décrit les résultats empiriques des quatre mois de données de la SOC qui ont motivé l'étude de simulation

¹ Katherine J. Thompson et Richard S. Sigman, Economic Statistical Methods and Programming Division, U.S. Bureau of the Census, Washington DC, 20233 U.S.A.

- HECKMAN, J.J., et BORJAS, G.J. (1980). Does unemployment cause future unemployment? Definitions, questions, and answers from a continuous time model of heterogeneous and state dependence. *Economica*, 47, 247-283.
- HESS, J., SINGER, E., et BUSHERY, J. (2000). Predicting test-retest reliability from behavior coding. *International Journal of Public Opinion Research*, II, 4, 346-360.
- HUI, S.L., et WALTER, S.D. (1980). Estimating the error rates of diagnostic tests. *Biometrics*, 36, 167-171.
- KLEINBAUM, D.G., KUPPER, L.T., et MULLER, K.E. (1988). *Applied Regression Analysis and Other Multivariate Methods*. Boston: PWS-KENT Publishing Co.
- LIU, T.H., et DAYTON, C.M. (1997). Model selection information criteria for non-nested latent class models. *Journal of Educational and Behavioral Statistics*, 22, 249 - 264.
- LYNCH, L.M. (1989). The youth labor market in the eighties: determinants of re-employment probabilities for young men and women. *The Review of Economics and Statistics*, 37-45.
- MEYERS, B. D. (1988). Classification-error models and labor-market dynamics. *Journal of Business and Economic Statistics*, 6, 3, 385-390.
- MOORE, J.C. (1988). Self-proxy response status and survey response quality. *Journal of Official Statistics*, 4, 2, 155-122.
- O'MURCHBARTAGH, C. (1991). Simple response Variance: Estimation and Determinants. *Measurement Errors in Surveys*, (P. Biemer, et al., Eds.). New York: John Wiley & Sons, 551-574.
- POTERBA, J., et SUMMERS, L. (1986). Reporting errors and labor market dynamics. *Econometrica*, 54, 6, 1319-1338.
- POTERBA, J., et SUMMERS, L. (1995). Unemployment benefits in classification. *The Review of Economics and Statistics*, 77, 207-216.
- POULSEN, C.S. (1982). Latent Structure Analysis with Choice Modeling Applications. Thèse de doctorat, Wharton School, University of Pennsylvania.
- ROTHGEB, J. (1994). Revisions to the CPS Questionnaire: Effects on Data Quality. U.S. Bureau of the Census. CPS Overlap Analysis Team Technical Report 2, April 6.
- VERMUNT, J. (1997). *ITEM: A General Program for the Analysis of Categorical Data*. Tilburg University.
- WIGGINS, L.M. (1973). *Panel Analysis, Latent Probability Models for Attitude and Behavior Processing*. Amsterdam: Elsevier S.P.C.
- SCHREINER, I. (1980). Rerinterview Results from the CPS Independent Reconciliation Experiment (Second Quarter 1978 through Third Quarter 1979). Internal U.S. Bureau of the Census Report.
- SHOCKEY, J. (1988). Adjusting for response error in panel surveys, a latent class approach. *Sociological Methods and Research*, 17, 1, 65-92.
- SINCLAIR, M., et GASTWIRTH, J. (1996). On procedures for evaluating the effectiveness of reinterview survey methods: application to labor force data. *Journal of the American Statistical Association*, 91, 961-969.
- SINCLAIR, M., et GASTWIRTH, J. (1998). Estimations des erreurs de classification dans l'enquête sur la population active et analyse de leur incidence sur les taux de chômage publics. *Techniques d'enquête*, 24, 2, 171-183.
- SINGH, A.C., et RAO, J.N.K. (1995). On the adjustment of gross flow estimates for classification error with application to data from the canadian labour force survey. *Journal of the American Statistical Association*, 90, 430, 478-488.
- U.S. BUREAU OF THE CENSUS (1985). Evaluating Censuses of Population and Housing. STD-ISP-TR-5. Washington, D.C.: U.S. Government Printing Office.
- U.S. BUREAU OF THE CENSUS (2000). Current Population Survey: Design and Methodology. U.S. Bureau of the Census Technical Paper 63. Washington, D.C.: Government Printing Office.
- VAN DE POL, F., et DE LEEUW, J. (1986). A latent markov model to correct for measurement error. *Sociological Methods and Research*, 15, 1-2, 118-141.
- VAN DE POL, F., et LANGHEHEINE, R. (1997). Separating change and measurement error in panel surveys with an application to labor market data. *Survey Measurement and Process Quality*, (L. Lyberg, et al., Eds.). New York: John Wiley & Sons.

réinterview ne sont typiquement pas disponibles dans les enquêtes par panel et, par conséquent, les analystes ne pourront peut-être appliquer que les critères (1), (2) et (5) ci-dessus pour vérifier la validité du modèle. L'hypothèse markovienne est la clé de la stratégie MLCA. Il se peut que des données recueillies au moyen d'un panel portent sérieusement atteinte à cette hypothèse. Heureusement, l'échec de l'hypothèse markovienne ne semble pas être un facteur important de la validité des estimations MLCA de l'erreur de classification de la population active de la CPS (voir le tableau 4).

BIBLIOGRAPHIE

- ABOWD, J., et ZELLMER, A. (1985). Estimating gross labor-force flows. *Journal of Business and Economic Statistics*, 3, 3, 254-283.
- AGRESTI, A. (1990). *Categorical Data Analysis*. New York: John Wiley & Sons.
- AKERLOF, G.A., et MAIN, G.M. (1980). Unemployment spells and unemployment experience. *The American Economic Review*, 70, 3, 885-893.
- BAILLAR, B. A. (1975). The effect of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-30.
- BIEMER, P., BUSHERY, J., et FLANAGAN, P. (1997). An Application of Latent Markov Models to the CPS. Internal U.S. Bureau of the Census Technical Report.
- BIEMER, P., et FORSMAN, G. (1992). On the quality of reinterview data with applications to the Current Population Survey. *Journal of the American Statistical Association*, 87, 420, 915-923.
- BOHRNSTEDT, G.W. (1983). Measurement. *Handbook of Survey Research*, (P.H. Rossi, R.A. Wright, et A.B. Anderson, Eds.). New York: Academic Press.
- BUSHERY, J., et KINDELBERGER, K. (1999). Simulation Examples for MLC Analysis. Internal U.S. Bureau of the Census Memorandum, Washington, DC, 70-122.
- CHUA, T.C., et FULLER, W.A. (1987). A model for multinomial response error applied to labor flows. *Journal of the American Statistical Association*, 82, 397, 46-51.
- CLOGG, C., et ELIASON, S. (1985). Some common problems in log-linear analysis. *Sociological Methods and Research*, 16, 8-14, 1-6.
- COHEN, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 210, 37-46.
- CORAK, M. (1993). Is unemployment insurance additive? Evidence from the benefit durations of repeat users. *Industrial and Labor Relations Review*, 47, 1, 62-72.
- FORSMAN, G., et SCHREINER, I. (1991). The design and analysis of reinterview: an overview. *Measurement Errors in Surveys*, (P.P. Biemer, et al., Eds.). New York: John Wiley & Sons. 279-302.
- FULLER, W., et CHUA, T.C. (1985). Cross change estimation in the presence of response error. *Proceedings of the Conference on Gross Flows in Labor Force Statistics*. Washington, D.C., U.S. Bureau of the Census and U.S. Bureau of Labor Statistics, 65-77.

1993, nous avons observé des différences faibles mais significatives entre les estimations MLCA et les estimations H-W correspondantes. Ces différences pourraient s'expliquer par des différences de période de temps, puisque les données de réinterview ont précédé de quelques années les données d'interview de la CPS.

4. Accord entre les estimations modélisées et de type

test-retest du manque de cohérence. Nous avons comparé des estimations MLCA modélisées de l'indice d'incohérence aux estimations directes correspondantes tirées du programme de réinterview de la CPS. Les deux ensembles d'estimations concordent assez bien pour les trois années, exception faite pour la catégorie inactif (voir le tableau 6). Pour 1995 et 1996, les différences s'expliquent en partie par le biais de l'estimateur traditionnel résultant de l'échec de l'hypothèse des mesures parallèles. La méthode H-W, qui n'exige pas l'hypothèse des mesures parallèles, permet de produire des estimations de l'indice qui concordent bien avec les estimations MLCA pour 1995 et 1996. Pour 1993, la différence entre les estimations MLCA et H-W est peut-être attribuable à des différences de période de temps pour les ensembles de données de réinterview et de la CPS.

5. Plausibilité des profils d'erreur de classification.

Les estimations MLCA des probabilités d'erreur de classification semblent plausibles. Les estimations d'un groupe avec/sans procuration à l'autre concordent avec les attentes antérieures voulant que des taux d'erreur moins élevés soient observés pour les répondants sans procuration que pour les répondants par procuration. De plus, les taux d'erreur les plus élevés ont été observés pour la population des sans-emploi et l'ordre de grandeur de ces estimations concordait avec celui d'études antérieures (voir par exemple Fuller et Chua 1985, Abowd et Zellner 1985, Poterba et Summers 1986, Sinclair et Gastwirth 1996; voir le tableau 3).

En résumé, nous n'avons trouvé dans ces analyses aucune raison de mettre en doute la validité de la stratégie MLCA. La méthode a donné de bons résultats pour les cinq épreuves de validité. Nous recommandons donc que la méthode MLCA soit considérée comme une autre méthode d'évaluation de l'exactitude des estimations de la population active de la CPS. La forte concordance entre les estimations MLCA et H-W confirme également la validité de la méthode H-W. Nous recommandons que les deux méthodes soient considérées dans de futures études de la qualité des données de la CPS.

Même si la stratégie MLCA a donné de bons résultats dans nos épreuves, nous recommandons que la méthode soit appliquée avec prudence à d'autres situations. Dans notre analyse, les données de réinterview ont permis d'évaluer la validité des estimations MLCA. Toutefois, les données de

sous-estime la fiabilité réelle des données de la CPS; l'indice d'incohérence de la CPS est donc trop élevé.

L'interview et la réinterview de la CPS comporteront des fiabilités différentes si la répartition des erreurs n'est pas égale pour les deux interviews. Il est possible de le vérifier en comparant l'ajustement d'un modèle de type H-W en présence et en l'absence de la restriction $\pi_{0|x} = \pi_{0|x'}$. L'hypothèse d'une fiabilité égale est rejetée si la différence entre les chi-carrés du rapport des vraisemblances pour les deux modèles est supérieure à un chi-carré ayant 6 degrés de liberté. Ce test a été rejeté pour 1995 et 1996 au niveau de signification de 10%. La différence entre les estimations de la catégorie inactif pour 1995 et 1996 semble donc attribuable en partie au biais des estimations traditionnelles de I.

Tableau 6

Méthode d'estimation	Classification de la population active		
	Employé	Sans emploi	Inactif

Comparaison des estimations MLCA, H-W et traditionnelles de l'indice d'incohérence selon l'année et la classification de la population active	Indice		
	active	inactif	global
1993	Estimation 8,16	33,49	9,96
	traditionnelle (0,24)	(1,16)	(0,26)
H-W	7,37	34,93	10,07
MLCA	6,35	28,04	7,63
1995	Estimation 6,69	36,28	10,80
	traditionnelle (0,44)	(2,85)	(0,53)
H-W	6,82	37	8,98
MLCA	6,06	36,19	7,2
1996	Estimation 5,93	35,97	11,95
	traditionnelle (0,39)	(2,68)	(0,56)
H-W	5,67	39,46	7,55
MLCA	5,99	37,39	7,76
			9,06

A noter également que les indices H-W et MLCA concordent assez bien pour 1995 et 1996, même s'ils diffèrent quelque peu en 1993. Toutefois, comme il a été noté à la discussion du tableau 5, les comparaisons entre les estimations MLCA et H-W pour cette année-là sont obscures par les différentes périodes de temps utilisées pour la construction de l'ensemble des données d'interview-réinterview antérieur à 1994. Cela pourrait expliquer au moins partiellement l'écart entre les estimations pour cette année-là.

La présente recherche a surtout tenté d'examiner la validité des estimations MLCA de l'erreur de classification de la population active de la CPS, et de déterminer l'efficacité de la stratégie MLCA comme substitut des méthodes

5. SOMMAIRE ET CONCLUSIONS

1. **Diagnostics modelisés.** Nous avons examiné tout un choix de modèles MLCA comportant des variables de regroupement définies selon l'âge, la race, le sexe, la scolarité, le mode d'interview et le type de réponse (avec ou sans procuration). Le modèle manifestant la plus grande parcimonie et le meilleur ajustement pour les trois années comporte une variable de regroupement définie par le type de réponse (avec ou sans procuration) selon quatre catégories: trois cycles tous menés sans procuration, deux cycles seulement menés par procuration et trois cycles tous menés par procuration. Pour cette catégorie de modèles, le meilleur a été le modèle 4 (voir le tableau 1), qui précisait des probabilités de changement non homogènes et non stationnaires et des probabilités de réponse non homogènes. Ce modèle a fourni un ajustement adéquat aux données pour les trois années.
2. **Validité de l'ajustement du modèle d'une année à l'autre de la CPS.** Un autre indicateur de la validité du modèle est son ajustement d'un échantillon indépendant à l'autre pour la même population. À supposer que la dynamique de la population active et que la structure des probabilités de réponse pour la CPS soient stables pendant quatre ans, le même modèle général devrait manifester un ajustement adéquat pour les trois années. Le modèle 4 affiche une validité de l'ajustement sur plusieurs années (voir le tableau 1). De plus, d'autres variables de regroupement ont été vérifiées dans l'étude, mais le modèle comportant la variable avec ou sans procuration a été le meilleur pour les trois années.
3. **Accord entre les estimations modelisées et de type test-retest des probabilités de réponse.** À l'aide des données d'interview-réinterview non rapprochées tirées de la CPS pour trois périodes (avant 1994, 1995, 1996), nous avons appliqué la méthode H-W à l'estimation des probabilités de réponse et nous avons comparé celles-ci aux estimations MLCA. Un bon accord a été observé pour 1995 et 1996, deux années pour lesquelles les périodes de temps pour les données de réinterview et les données de la CPS manifestaient une étroite concordance (voir le tableau 5). Pour

En général, les deux ensembles d'estimations sont en assez bon accord. Les années 1995 et 1996 ne manifestent aucune différence statistiquement significative (au niveau de 5%) entre les estimations MLCA et H-W pour ce qui est de la population des sans-emploi. Les estimations antérieures à 1994 affichent des différences appréciables; toutefois, celles-ci s'expliquent peut-être par le fait que les données de réinterview antérieures à 1994 ont été recueillies de 1985 à 1988, plutôt qu'en 1993. Ces différences sont étudiées plus en profondeur dans la section qui suit.

4.2 Comparaison des indices d'incohérence

Comme il a été décrit à la section 3.1, nous estimons l'indice d'incohérence pour chaque période de temps à l'aide des estimations MLCA modélisées des probabilités de réponse. Essentiellement, nous estimons le tableau de croisé prévu des interviews-réinterviews à partir des estimations MLCA des probabilités de réponse, et nous appliquons ensuite la formule pour l'indice à ce tableau comme si le tableau était observé. Un deuxième tableau de classification prévu des interviews-réinterviews peut être estimé à l'aide des estimations H-W des probabilités de réponse. Nous comparons ensuite ces deux ensembles d'estimations à l'estimation de l'indice obtenue directement des données de réinterview de la CPS à l'aide de méthodes traditionnelles (U.S. Bureau of the Census 1985). L'accord entre les trois estimations confirme la validité des trois méthodes.

Le tableau 6 montre les trois méthodes d'estimation de l'indice d'incohérence pour les trois périodes de temps. Comme auparavant, la catégorie sans emploi revêt un intérêt particulier à cause de son taux d'erreur élevé. Puisque les erreurs-types ne sont pas disponibles pour les estimations MLCA ou H-W de l'indice, il n'est pas possible de vérifier l'hypothèse formellement. Toutefois, il existe des erreurs-types pour les estimations traditionnelles que l'on peut utiliser comme approximation des erreurs-types pour les estimations H-W et traditionnelles pour les trois années. MLCA sont en assez bon accord généralement avec les estimations H-W et traditionnelles pour les trois années. Toutefois, pour ce qui est de la catégorie inactif en 1996 et en 1996, les estimations traditionnelles de l'indice sont un peu plus grandes que l'une ou l'autre des estimations modélisées de structure latente. Une analyse plus poussée indique que cette différence est attribuable à un biais de la stratégie d'estimation traditionnelle résultant de l'invalidité de l'hypothèse des mesures parallèles.

Le Bureau of the Census des États-Unis (1985) a montré que si les processus d'interview et de réinterview comportent des fiabilités différentes, l'estimation traditionnelle de l'indice est biaisée. Si, par exemple, la fiabilité des données de réinterview est inférieure à celle des données d'interview, l'estimateur traditionnel de la fiabilité test-rétest

hommes blancs et des femmes blanches et à deux catégories d'activité: inactif et actif. Cette dernière catégorie est la somme de nos catégories employé et sans emploi. Dans notre analyse, nous considérons des membres d'échantillon de toutes les races et nous analysons la classification à trois catégories de la situation vis-à-vis de l'activité utilisée dans la stratégie MLCA. Ainsi, l'analyse H-W estime 16 paramètres pour chaque année, ce qui correspond au nombre de degrés de liberté disponibles du tableau $G \times A \times A$, ne laissant aucun degré de liberté pour vérifier l'ajustement du modèle. Le logiciel IBM a été utilisé pour l'ajustement du modèle H-W aux données d'interview et de réinterview non rapprochées pour trois périodes de temps coïncidant avec les trois périodes de notre stratégie MLCA: avant 1994, 1995, 1996. Nous avons tenté de limiter l'analyse au premier trimestre seulement de ces périodes de temps. Malheureusement, à cause de la faible taille des échantillons, les estimations ont été plutôt instables. Il a donc fallu utiliser les données de réinterview de l'ensemble des quatre trimestres de ces périodes de temps. Les données antérieures à 1994 ont été recueillies de 1985 à 1988 grâce à un échantillon de réinterviews non rapprochées.

Les résultats de cette comparaison des estimations MLCA et H-W sont résumés dans le tableau 5. Les estimations MLCA sont les mêmes que celles des lignes du tableau 2 étiquetées «Total». Les estimations H-W sont les probabilités de classification associées à l'interview originale, c'est-à-dire la mesure λ en (19). Le tableau montre la comparaison pour les trois années. Puisque le taux d'erreur le plus élevé de la stratégie MLCA est survenu pour la catégorie sans emploi, celle-ci revêt un intérêt particulier dans la comparaison MLCA/H-W.

Tableau 5

Comparaison d'estimations modélisées MLCA et H-W des probabilités de réponse de la CPS selon l'année (écarts-types entre parenthèses)

Classification	Réelle	Employé	Sans emploi	1993						1995						1996					
				MLCA	H-W	MLCA	H-W	MLCA	H-W	MLCA	H-W	MLCA	H-W	MLCA	H-W	MLCA	H-W	MLCA	H-W	MLCA	H-W
Employé	Sans emploi	0,0	0,0	(0,3)	(0,1)	0,0	0,0	(0,3)	(0,1)	0,0	0,0	(0,3)	(0,1)	0,0	0,0	(0,3)	(0,1)	0,0	0,0	(0,3)	(0,1)
	Inactif	0,7	0,9	(0,1)	(0,0)	0,7	0,9	(0,1)	(0,0)	0,7	0,9	(0,1)	(0,0)	0,7	0,9	(0,1)	(0,0)	0,7	0,9	(0,1)	(0,0)
	Sans emploi	11,1	7,1	(0,3)	(1,0)	11,5	7,9	(0,7)	(2,3)	12,5	8,6	(0,1)	(1,5)	13,1	9,1	(0,1)	(1,5)	13,6	10,1	(0,1)	(1,5)
Sans emploi	Sans emploi	74,3	81,8	(2,7)	(1,1)	76,1	87,6	(2,9)	(1,3)	77,4	91,1	(2,9)	(1,3)	78,8	94,6	(2,9)	(1,3)	80,2	97,4	(2,9)	(1,3)
	Inactif	14,7	11,1	(2,9)	(0,9)	15,6	12,6	(3,0)	(1,2)	16,6	14,1	(3,0)	(1,2)	17,6	15,6	(3,0)	(1,2)	18,6	17,6	(3,0)	(1,2)
	Sans emploi	2,0	1,4	(0,5)	(0,1)	2,5	1,1	(0,5)	(0,1)	2,6	1,1	(0,5)	(0,1)	2,7	1,1	(0,5)	(0,1)	2,8	1,1	(0,5)	(0,1)
Inactif	Sans emploi	96,8	97,8	(0,6)	(0,1)	97,0	98,2	(0,6)	(0,1)	97,0	98,2	(0,6)	(0,1)	97,0	98,2	(0,6)	(0,1)	97,0	98,2	(0,6)	(0,1)
	Inactif	1,2	0,8	(0,3)	(0,1)	1,2	0,8	(0,3)	(0,1)	1,2	0,8	(0,3)	(0,1)	1,2	0,8	(0,3)	(0,1)	1,2	0,8	(0,3)	(0,1)
	Sans emploi	96,8	97,8	(0,6)	(0,1)	97,0	98,2	(0,6)	(0,1)	97,0	98,2	(0,6)	(0,1)	97,0	98,2	(0,6)	(0,1)	97,0	98,2	(0,6)	(0,1)

nous avons considéré des structures latentes comportant uniquement deux classes ou situations à un moment donné: chômage, noté X_i , Y ou $Z = 1$, et autre (employé ou inactif), noté X_i , Y ou $Z = 2$ avec des définitions semblables pour les situations observées A_i , B et C . Afin d'établir une population en vue de la simulation, les probabilités latentes $\pi_{a|x}$, $\pi_{y|x}$ et $\pi_{z|x}$ ont été précisées comme concordant avec les ensembles combinés de données de 1993, 1995 et 1996. Nous avons alors défini deux paramètres, λ_1 et λ_2 , variant dans la simulation, où

$$\lambda_1 = \pi_{z=1|x=2,y=1} = \pi_{z=1|x=1,y=1} \quad (17)$$

$$\lambda_2 = \frac{\pi_{z=1|x=2,y=2}}{\pi_{z=1|x=1,y=2}} = \frac{\pi_{z=2|x=1,y=2}}{\pi_{z=2|x=2,y=2}} \quad (18)$$

Ainsi, λ_1 est la probabilité qu'une personne soit «sans emploi» en mars, après avoir été «sans emploi» en février et «autre» en janvier relativement à la probabilité d'être «sans emploi» en mars après avoir été «sans emploi» au cours des deux mois précédents. En conformité avec les résultats obtenus par Akterlof et Main (1980), qui ont montré que la probabilité qu'une personne demeure sans emploi augmente à mesure que le nombre de situations sans emploi augmente, nous supposons que $0 < \lambda_1 \leq 1$. De même, λ_2 est la probabilité qu'une personne soit «sans emploi» en mars, après avoir été «sans emploi» en février et «autre» en janvier relativement à la probabilité d'être «sans emploi» après avoir été «autre» en février et «sans emploi» en janvier. Encore une fois, suivant Akterlof et Main, nous supposons que $0 < \lambda_2 \leq 1$. À noter que, lorsque $\lambda_1 = \lambda_2 = 1$, les changements de chômage entre février et mars sont markoviens.

Les données simulées ont été produites de façon à correspondre complètement à un modèle MLCA à probabilités de changement non stationnaires lorsque $\lambda_1 = \lambda_2 = 1$. Nous avons simulé l'échec de l'hypothèse markovienne en variant λ_1 et λ_2 entre 0 et 1. Par souci d'uniformité avec les données 1993-1996, nous avons fixé la probabilité d'une réponse «sans emploi» correcte, $\pi_{a=1|x=1}^a$ à 0,80 et la probabilité d'une réponse «autre» correcte, $\pi_{a=2|x=2}^a$ à 0,99 dans toutes les simulations. De plus, les dénominateurs de λ_1 et λ_2 ont été fixés à leurs valeurs déterminées à partir des données combinées 1993-1996, tandis que les numérateurs ont été calculés à partir de (17) et de (18) à l'aide de valeurs de λ_1 et de λ_2 précisées pour chaque exécution d'une simulation.

Le tableau 4 résume les résultats de la simulation pour $\lambda_1 = \lambda_2 = \lambda$, où l'on fait varier λ entre 0,2 et 1,0 par paliers de 0,2. À noter que pour $\lambda_1 = \lambda_2 = 1,0$, qui correspond à un modèle markovien, les probabilités estimatives de réponse correcte sont exactement comme prévu. Pour des valeurs plus petites de λ_1 et de λ_2 , les estimations manifestent un biais négatif, le biais maximal correspondant à la valeur la plus faible envisagée, 0,2. Néanmoins, le biais absolu attribuable à des probabilités de changement non markovien ne

dépasse jamais 3 points de pourcentage. Les résultats du tableau 4 correspondent à ceux de Bushery et Kindelberger (1999), qui ont utilisé une stratégie un peu différente afin d'illustrer le même caractère de robustesse des modèles MLCA pour des données de la CPS. Les deux études indiquent que le manque de validité de l'hypothèse markovienne ne semble pas être une source importante de biais dans l'estimation des probabilités d'erreur de classification de la CPS.

Tableau 4

Estimations de classification correcte pour des changements non markoviens

(les entrées de cellules sont des pourcentages)

Pr (correct)	$\lambda_1 = \lambda_2 = \lambda$			
	$\lambda = 0,2$	$\lambda = 0,4$	$\lambda = 0,6$	$\lambda = 0,8$
(Markov)	800	787	793	800
Pr («chôm.» réel)	776	781	787	988
Pr («autre» réel)	986	987	988	990
$\pi^a = \pi^{a=1 x=1} = \pi^{a=2 x=2}$				

4. COMPARAISON DES ESTIMATIONS DE TYPE MLC A ET DE RÉINTERVIEW NON RAPPROCHÉE

4.1 Estimation de Hui-Walter

On peut obtenir un autre ensemble d'estimations des probabilités de réponse des données de réinterview de la CPS à l'aide d'un type de modèle de structure latente d'abord proposé par Hui et Walter (1980). Nous utilisons la notation mentionnée ci-dessus; soit X_i , la classification réelle de la situation vis-à-vis de l'activité à un moment donné; soit A et A' , les classifications d'interview et de réinterview, respectivement. Soit G , une variable de groupe ment définie comme en (4). Considérons la vraisemblance du tableau groupex interviewx réinterview noté $G \times A \times A'$. La probabilité qu'une personne relevant du groupe G soit classée dans la cellule (a, a') du tableau. Le modèle pour $\pi_{a a'}^{g a'}$ proposé par Hui et Walter est le suivant:

$$\pi_{a a'}^{g a'} = \sum_{x=1}^x \pi_{a|x}^g \pi_{a'|x}^g \pi_{a|x}^a \pi_{a'|x}^a \quad (19)$$

Dans ce modèle, l'hypothèse des mesures parallèles pour les réponses d'interview et de réinterview est réduite et les probabilités de réponse pour les deux mesures, c'est-à-dire $\pi_{a|x}^a$ et $\pi_{a'|x}^a$, sont estimées séparément. L'hypothèse ICF est adoptée comme condition de l'identifiabilité. Il est également supposé que $\pi_{a|x}^a$ et $\pi_{a'|x}^a$, ne dépendent pas de la variable de groupe, G , tandis que $\pi_{a|x}^g$ et $\pi_{a'|x}^g$, dépendent des groupes employés, sans emploi et inactif, c'est-à-dire $\pi_{x|g}$, de- pend toujours de G . Sinclair et Gastwirth (1996), dans leur analyse de l'erreur de classification de la population active de la CPS, ont utilisé le sexe comme variable de groupement, et nous utilisons également cette variable de groupement dans notre analyse. Sinclair et Gastwirth ont limité leur analyse à des

Tableau 3

Comparaison des estimations MLCa aux estimations publiées antérieurement																															
Classification		MLCA		Chua et Fuller (données de 1982)		Poterba et Summers (données de 1981)		CPS		Réinteriew (1977-1982)		Réelle		Observée																	
Employé	Employé	98,77 (1993)	98,66 (1 mois)	98,65 (2 mois)	98,73 (1995)	98,73 (1996)	0,34 (1993)	0,49 (1995)	0,37 (1996)	0,89 (1993)	0,78 (1995)	0,79 (1996)	7,06 (1993)	7,86 (1995)	8,57 (1996)	81,81 (1993)	76,09 (1995)	74,42 (1996)	11,13 (1993)	16,04 (1995)	17,00 (1996)	1,41 (1993)	1,11 (1995)	1,13 (1996)	0,75 (1993)	0,69 (1995)	0,87 (1996)	97,84 (1993)	98,20 (1995)	98,00 (1996)	
Sans emploi	Sans emploi	98,73 (1996)	98,65 (2 mois)				0,32 (1 mois)	0,34 (2 mois)		1,02 (1 mois)	1,01 (2 mois)		3,52 (1 mois)	3,51 (2 mois)		88,27 (1 mois)	88,23 (2 mois)		8,21 (1 mois)	8,16 (2 mois)		1,60 (1 mois)	1,61 (2 mois)		1,19 (1 mois)	1,24 (2 mois)		97,15 (2 mois)	97,12 (1 mois)		
						</																									

Une explication de cette différence est que les estimations comparatives manifestent un biais vers le haut à cause de corrélations entre les erreurs d'interview et de réinterview. Une autre explication est que les estimations MLCa de l'hypothèse markovienne. Nous estimons que les deux explications sont valables dans une certaine mesure. Toutefois, nous indiquons à la section suivante que l'échec de l'hypothèse markovienne exerce probablement une faible influence sur les estimations de l'erreur de classification.

3.6 Robustesse de la stratégie MLCa à l'égard des changements non markoviens de la situation vis-à-vis de l'activité

Plusieurs auteurs ont examiné l'effet de la situation actuelle et antérieure vis-à-vis de l'emploi sur la situation future vis-à-vis de l'emploi (voir par exemple Akkerlof et Main 1980; Heckman et Borjas 1980; Lynch 1989; Corak 1993). Heckman et Borjas ont indiqué que l'étude de cette question est assez difficile à cause des biais de sélection, des erreurs de réponse et de l'hétérogénéité non observée. Ces sources de confusion rendent peut-être compte des résultats

Afin de pouvoir étudier l'effet des infractions à l'hypothèse markovienne en (2) pour la présente application, nous avons mené une étude de simulation limitée. Afin d'orienter la recherche tout en simplifiant le cadre de la simulation, les estimations MLCa de l'erreur de classification.

peu cohérents de la documentation. Ainsi, à l'aide de données de la CPS, Akkerlof et Main (1980) ont pu indiquer que la probabilité d'un futur chômage dépend du nombre de situations antérieures de chômage de même que de la durée de ces situations. Toutefois, dans une étude d'hommes ayant terminée leurs études secondaires, Heckman et Borjas (1980) n'ont trouvé aucune indication que les situations antérieures de chômage ou leur durée ont un effet sur le futur comportement vis-à-vis de l'activité lorsque l'on tient compte du biais de tirage de l'échantillon et du biais dû à l'hétérogénéité. Les résultats tirés de la documentation sont également incohérents et ambigus relativement à la mesure dans laquelle l'hypothèse markovienne exprimée en (2) risque d'être enfreinte pour la CPS et d'autres enquêtes sur la population active. Néanmoins, dans la présente section, nous cherchons à fournir une réponse au moins partielle à la question de savoir comment les changements non markoviens de la situation vis-à-vis de l'activité influencent les estimations MLCa de l'erreur de classification.

Tableau 1
Diagnostiques modelisés pour d'autres modèles MLCA selon l'année

Données de 1993						
Modèle de base: Changements homogènes et stationnaires et probabilités de réponse						
90	17	645	0,000	-320	0,048	
84	23	632	0,000	-269	0,047	
66	41	99	0,006	-609	0,007	
42	65	64	0,016	-386	0,005	
Modèle 3: Changements non homogènes et non stationnaires						
24	83	23	0,501	-234	0,002	
Modèle 4: Changements non homogènes et non stationnaires et probabilités de réponse non homogènes						
Données de 1995						
<i>df</i>	<i>npai</i> ¹	<i>L</i> ²	valeur <i>p</i>	BIC	<i>d</i>	
Modèle de base: Changements homogènes et stationnaires et probabilités de réponse						
90	17	697	0,000	-275	0,044	
84	23	668	0,000	-240	0,043	
66	41	146	0,000	-567	0,008	
42	65	82	0,000	-372	0,005	
Modèle 3: Changements non homogènes et non stationnaires						
24	83	25	0,410	-234	0,002	
Modèle 4: Changements non homogènes et non stationnaires et probabilités de réponse non homogènes						
Données de 1996						
<i>df</i>	<i>npai</i> ¹	<i>L</i> ²	valeur <i>p</i>	BIC	<i>d</i>	
Modèle de base: Changements homogènes et stationnaires et probabilités de réponse						
90	17	632	0,000	-325	0,045	
84	23	585	0,000	-308	0,044	
66	41	159	0,000	-543	0,010	
42	65	82,6	0,000	-364	0,005	
Modèle 3: Changements non homogènes et non stationnaires						
24	83	39,3	0,026	-216	0,003	
Modèle 4: Changements non homogènes et non stationnaires et probabilités de réponse non homogènes						
1 A noter que «npai» représente le nombre de paramètres du modèle.						

pour tous les p . En vertu de ces contraintes, (4) peut s'écrire comme suit

$$\pi_{d,a,b,c} = \sum_{x,y,z} \pi_{d,x|p}^d \pi_{y|xp}^d \pi_{z|py}^d (\pi_{a|p,x}^a)^3.$$

Dans le tableau 1, nous indiquons la statistique d'ajustement de base pour l'ensemble des cinq modèles et des trois années. La colonne 4 du tableau indique la statistique chi-carré habituelle du rapport des vraisemblances L^2 (voir Agresti 1990, 48), tandis que la colonne 5 indique la valeur p correspondante. Une valeur p égale ou supérieure à 0,05 est le critère habituel d'un ajustement de modèle adéquat. Toutefois, à cause de la taille élevée des échantillons de notre analyse, le fait d'exiger une valeur p aussi grande pourrait entraîner un surajustement du modèle. Nous considérons une valeur p aussi petite que 0,01 comme étant acceptable. La mesure BIC du tableau se définit comme suit:

$$BIC = L^2 - (\log N) df$$

où N est la taille de l'échantillon total et df représente les degrés de liberté du modèle. Le BIC résume essentiellement les compromis entre l'ajustement du modèle (L^2) et la parcimonie du modèle (df). Puisque de petites valeurs du BIC sont favorables, nous considérons le modèle comportant le plus faible BIC comme le meilleur relativement à la validité de l'ajustement et à la parcimonie. Liu et Dayton (1997) ont discuté de cette stratégie pour des modèles de structure latente.

3.5 Estimation de l'erreur de classification

Enfin, l'indice de dissimilitude (d) est la proportion des observations qui auraient à changer de cellule pour que le modèle soit parfaitement ajusté. En général, on considère des modèles comportant $d \leq 0,05$ (erreur de modèle de 5%) comme étant bien ajustés aux données (Vermunt 1997). Pour chaque année de données, le modèle 4 est le seul à fournir un ajustement acceptable lorsque le critère des valeurs p est considéré. Le modèle 4 est également plausible du point de vue de la théorie des réponses puisqu'il postule que l'erreur de classification varie selon le groupe Self/Proxy. Cela, comme nous l'avons dit, correspond à la documentation publiée sur les méthodes d'enquête (voir par exemple O'Muircheartaigh 1991 et Moore 1988). L'indice de dissimilitude d , pour le modèle est de 0,3%, ce qui indique un ajustement de modèle très valable. Nous utilisons donc le modèle 4 pour produire les estimations de l'erreur de classification de la population active.

Le tableau 2 montre les estimations des probabilités de réponse tirées du modèle 4 pour les catégories employé, sans emploi et inactif. Pour les personnes réellement employées et celles réellement inactives, la probabilité d'une réponse correcte est assez élevée: au moins 98% pour les employés et 97% pour les inactifs. Toutefois, pour les personnes réellement sans emploi, la probabilité d'une réponse correcte varie d'une année à l'autre et d'un groupe à l'autre entre 68% environ et 86% environ. Comme prévu,

dans le groupe Proxy est légèrement plus élevé (un tiers au lieu d'un quart).

3.4 Ajustement des modèles MLCA

Pour l'ajustement d'un modèle MLCA comportant une seule variable de groupement, P , l'ensemble de données d'entrée était un tableau $4 \times 3 \times 3 \times 3$ à dénombrement de cellules défini par le tri croisé de $P \times A \times B \times C$, où A , B et C représentent la classification de la situation vis-à-vis de l'activité pour janvier, février et mars, respectivement.

Le logiciel EBM et d'autres logiciels d'ajustement de modèle MLCA supposent un échantillonnage aléatoire simple, de sorte que le plan de sondage complexe de la CPS ne peut pas être modélisé exactement. Il est possible de rendre compte de la structure d'échantillonnage à probabilités inégales de la CPS à l'aide de dénombrements de cellules pondérés et remis à l'échelle plutôt que de totaux de cellules bruts (Clogg et Eliason 1985). Toutefois, le recours à des données non pondérées pour la stratégie MLCA offre deux avantages importants. Premièrement, nous pouvons comparer les estimations MLCA aux estimations tirées des études citées antérieurement de l'erreur de classification de la CPS, toutes fondées sur des données non pondérées.

Deuxièmement, les données de réinterview de la CPS servent à évaluer les critères 3 et 4 ne sont pas pondérées et aucun poids n'est disponible. Par conséquent, une partie au moins de l'analyse exige des données non pondérées; l'utilisation de données pondérées pour les autres critères risquerait de produire de fausses incohérences dans les résultats.

Afin d'examiner la validité d'inférences pour la population totale fondée sur une analyse non pondérée est l'estimation des erreurs-types. Puisqu'elles sont calculées en fonction d'hypothèses d'échantillonnage aléatoire simple, les estimations EBM de l'erreur-type risquent d'être sous-estimées par manque de connaissance de l'effet de groupement dans l'échantillon de la CPS. Pour tenir compte approximativement de ce phénomène, nous pouvons multiplier les variances EBM par un effet de plan de sondage calculé à partir des estimations de la population active de la CPS. Le Bureau of the Census des États-Unis (2000, 14-9) a indiqué que les effets de plan de sondage pour les estimations de la population active de la CPS ne dépassent pas 1,3; le fait de multiplier les erreurs-types EBM par $(1,3)^{1/2}$ devrait entraîner une inflation suffisante des erreurs-types pour tenir compte du groupement. Une stratégie équivalente

consiste à utiliser un niveau de signification de 3% plutôt que de 5% au moment de déclarer la différence entre deux estimations. Cette dernière stratégie sera considérée comme appropriée dans l'analyse à venir. Nous croyons que cela représente un test prudent puisque l'effet du plan de sondage de la CPS reflète l'augmentation de la variance attribuable tant au groupement de l'échantillon qu'à la pondération inégale, tandis que seuls des effets de groupement figurent dans nos estimations non pondérées.

Le tableau 1 indique les résultats de l'ajustement d'une série de modèles MLCA de plus en plus complexes pour chacun des trois ensembles de données. Le modèle de base est le modèle MLCA le plus simple; il précise que les probabilités de changement et les probabilités de réponse sont homogènes (c'est-à-dire qu'elles ne diffèrent pas selon le groupe P) et stationnaires (c'est-à-dire qu'elles sont identiques pour les trois mois). Ce modèle peut s'écrire comme suit

$$(11) \quad \pi_{p,a,b,c} = \sum_{x,y,z} \pi_{p,x} \pi_{x,y} \pi_{y,z} \pi_{z,p}$$

que l'on obtient de (4) en imposant les contraintes

$$(12) \quad \pi_{z|yp} = \pi_{y|xp} = \pi_{y|ix}$$

et

$$(13) \quad \pi_{a|xp} = \pi_{b|yp} = \pi_{c|zp} = \pi_{a|ix} = \pi_{b|iy} = \pi_{c|iz}$$

pour tous les p .

Pour le modèle 1 nous réduisons la contrainte (12) à

$$(14) \quad \pi_{z|yp} = \pi_{y|xp} \text{ pour } p = 1, \dots, 4$$

ce qui permet aux changements survenus de janvier à février et de février à mars de varier selon le groupe Self/Proxy, P . Pour le modèle 2, nous réduisons la contrainte (12) davantage à

$$(15) \quad \pi_{y|xp} = \pi_{y|ix} \text{ et } \pi_{z|yp} = \pi_{z|iy}$$

pour tous les p . Le modèle 3 réduit les contraintes tant d'homogénéité que de stationnarité pour les probabilités de changement de sorte que $\pi_{y|xp} \neq \pi_{z|yp}$. Ce modèle permet donc aux probabilités de changement de varier selon le groupe et selon le mois. Toutefois, les probabilités de réponse doivent toujours être égales d'un groupe à l'autre et d'un mois à l'autre.

Le modèle 4 est le plus général des modèles identifiables que nous avons considérés. Le modèle 4 permet aux probabilités de changement de janvier à février et de février à mars de varier indépendamment de l'un à l'autre des quatre groupes Proxy/Self. Ce modèle précise également que les probabilités de réponse sont les mêmes pour janvier, février et mars, mais qu'elles peuvent varier de l'un à l'autre des quatre groupes Proxy/Self. Nous avons obtenu ce modèle à partir du modèle 3 en réduisant les contraintes exigeant des

$$(16) \quad \pi_{a|xp} = \pi_{b|yp} = \pi_{c|zp}$$

cours d'une deuxième série de quatre mois consécutifs. La MLCA exige au moins trois interviews consécutives pour l'identifiabilité des paramètres de modèle. Nous avons un choix d'ensembles de données comprenant toutes les personnes interviewées au cours de trois ou quatre mois consécutifs de la CPS. Puisque le recours à quatre mois de données réduirait la taille de l'échantillon pour l'analyse de moitié, nous avons choisi de faire porter l'analyse sur trois mois consécutifs (janvier, février et mars) pour les trois années de données. Les cas de non-réponse et les cas de changement du ménage tout entier pour un ou plusieurs des trois mois ont été exclus de l'analyse.

Le modèle MLCA le plus simple précise que les probabilités de réponse π_{a1x} , π_{b1y} , et π_{c1z} , ainsi que les probabilités de changement, π_{yx} , π_{xy} , sont identiques pour toutes les personnes de la population cible (c'est ce qu'on appelle l'homogénéité). Toutefois, notre analyse préliminaire (Biemer, Bushery et Flanagan 1997) a indiqué que les probabilités de réponse et de changement ne sont pas homogènes. Afin de rendre compte de cette hétérogénéité, nous avons examiné un certain nombre de covariables et de variables de stratification à inclure dans les modèles, y compris: le sexe, la scolarité, le mode d'interview, le type de réponse (avec ou sans procuration) et la race. Parmi celles-ci, c'est la variable tirée de l'indicateur de réponse avec ou sans procuration de la CPS qui a le mieux tenu compte de l'hétérogénéité de la population. Cette variable, notée P , se définit comme suit:

$$P = \begin{cases} 1 & \text{si les trois interviews sont toutes menées sans procuration (SELF)} \\ 2 & \text{si deux des trois interviews sont menées sans procuration (MOSTLY SELF)} \\ 3 & \text{si deux des trois interviews sont menées par procuration (MOSTLY PROXY)} \\ 4 & \text{si les trois interviews sont toutes menées par procuration (PROXY)} \end{cases}$$

À noter que nous utilisons maintenant P pour représenter la variable de sectionnement, au lieu de G , que nous avons utilisé à la section 2. Compte tenu de recherches antérieures (par exemple O'Muircheartaigh 1991), nous nous attendons à ce que le groupe Self ($P=1$) comporte moins d'erreurs de classification que le groupe Proxy ($P=4$). Nous vérifions cette hypothèse dans le cadre du critère de plausibilité des estimations (critère 4 ci-dessus).

La taille des échantillons pour les trois ensembles de données utilisés dans notre analyse est la suivante:

1993:	45 291 personnes
1995:	49 347 personnes
1996:	41 751 personnes

Pour 1993, un tiers environ de l'échantillon se trouve dans le groupe Self, un quart environ dans le groupe Proxy, les autres membres de l'échantillon étant répartis à peu près également entre les groupes Mostly Self et Mostly Proxy. Pour 1995 et 1996, le nombre de membres de l'échantillon

4. Accord entre les estimations modélisées et

test-*retest* de l'indice d'incohérence. Ce critère est semblable au troisième critère car il compare des estimations tirées de la MLCA à des estimations fondées sur des données de réinterview non rapprochées. Toutefois, cette analyse ne fait pas appel à la validité de la méthode d'estimation de Hui-Walter pour l'évaluation de la validité des estimations MLCA. Nous utilisons plutôt les estimations MLCA de l'erreur de classification pour obtenir des estimations de l'indice d'incohérence à l'aide de (7) à (9). Nous comparons ces estimations de la fiabilité directe-ment aux estimations de la fiabilité tirées du programme de réinterview de la CPS, obtenues de données de réinterviews non rapprochées. Un bon accord entre les estimations de réinterview et de type MLCA confirme la validité des deux méthodes, tandis qu'un faible accord indique qu'une des stratégies au moins n'est pas valable.

5. Plausibilité des profils d'erreur de classification.

Enfin, la plausibilité (ou la validité apparente) des estimations des probabilités de réponse peut également servir d'épreuve de la validité. Par exemple, il semble peu plausible que des réponses par procuration exactes que des réponses sans procuration. D'autres profils d'erreur de classification peuvent être examinés et évalués en fonction de la plausibilité. Pour autant que les estimations modélisées semblent plausibles, leur validité apparente est confirmée.

À la section suivante, nous discutons de nos résultats de modélisation MLCA dans le contexte de ces critères de validité. Nous commençons par décrire les ensembles de données de la CPS et les résultats du processus de sélection du modèle.

3.3 Ensembles de données de la CPS

En 1994, dans le cadre de la mise en œuvre de l'IPAO (interview sur place assistée par ordinateur), la CPS a été remaniée de fond en comble, avec restructuring des questions servant à déterminer la situation vis-à-vis de l'activité. Rothgeb (1994) a décrit le remaniement de la CPS. Comme conséquence de ces améliorations, nous nous attendons à observer une différence (plus précisément une réduction) de l'erreur de classification de la CPS après 1994 comparative-ment à 1993. Bien qu'il ne s'agisse pas d'un objectif principal de la présente recherche, nous avons comparé l'erreur de la CPS avant et après le remaniement. Nous avons vérifié la stratégie MLCA pour trois années de la CPS: 1993, 1995 et 1996, puisque des données de réinterview non rapprochées de la CPS étaient facilement accessibles pour cette période de temps.

Les ménages de la CPS sont interviewés au cours de quatre mois consécutifs, ne font pas partie de l'enquête pendant huit mois, puis sont de nouveau interviewés au

1. Diagnostics modelisés. Une condition nécessaire de la validité d'un modèle est que celui-ci soit plausible (que les hypothèses soient raisonnables et concordent avec la réalité) et qu'il ajuste les données de façon adéquate. Nous utilisons le critère traditionnel chi-carré de la validité de l'ajustement ainsi que d'autres mesures diagnostiques de l'ajustement du modèle afin d'évaluer le bien-fondé de celui-ci et la mesure dans laquelle les données concordent avec lui.

2. Validité de l'ajustement du modèle d'une année à l'autre de la CPS. Une méthode répandue de validation d'un modèle consiste à évaluer l'ajustement du modèle pour des données qui sont indépendantes des données utilisées pour la modélisation (voir par exemple Kleinbaum, Kupper et Muller 1988, 330). Cette méthode permet d'éviter un surparamétrage du modèle et une sélection de modèle guidée par les données (plutôt que par la théorie). Dans la présente étude, l'ajustement du même modèle aux données de chaque année séparément est un aspect de cette technique de validation indépendante du modèle. L'accord du modèle d'une année à l'autre aurait tendance à confirmer la validité de sa structure. Cette méthode comporte une difficulté pour la présente application. Après 1993, le questionnaire imprimé de la CPS a été remanié en fonction de l'PAO (interview sur place assistée par ordinateur), de sorte que l'ampleur des erreurs de réponses a pu changer après 1993. Toutefois, si les principales sources d'erreur de réponse de la CPS n'ont pas changé sous l'effet du remaniement, une structure modélisée qui décrit convenablement l'erreur pour 1993 devrait également décrire l'erreur pour 1995 et 1996.

3. Accord entre les estimations MLCA et les estimations test-retest de Hui-Walter (Hui et Walter 1980) d'estimation des probabilités de réponse. La méthode de Hui-Walter (1996, 1998). Même si les méthodes MLCA et H-W font toutes deux appel à des modèles de structure latente, les hypothèses de modélisation sont très différentes. Par exemple, la méthode H-W n'exige pas de données pour la méthode H-W sont indépendantes de celles utilisées pour la méthode MLCA. Le bon accord entre les deux ensembles d'estimations confirme la validité des deux méthodes, tandis qu'un faible accord indique qu'une des stratégies au moins n'est pas valable. Une forte concordance entre les estimations MLCA et H-W fournit également une certaine assurance que les estimations MLCA des probabilités de réponse sont relativement robustes vis-à-vis d'infractious éventuelles à l'hypothèse markovienne.

Comme nous l'avons expliqué à la section précédente, la MLCA des données longitudinales de la CPS fournira des estimations du maximum de vraisemblance de $\pi^{A_{1X}}$ et de π^X , permettant l'estimation de π^{A^A} à l'aide de (6) et de (7). Nous pouvons estimer la fiabilité test-retest, R , pour toute situation vis-à-vis de l'activité en appliquant les méthodes d'estimation habituelles (voir par exemple Bohmstedt 1983) à cette estimation de π^{A^A} . Pour notre analyse, nous calculons l'indice d'incohérence, $I = 1 - R$, qui représente la mesure traditionnelle de la fiabilité des données sur la population active de la CPS (voir U.S. Bureau of the Census 1985). Soit I_a , l'indice d'incohérence pour la catégorie $A = a$. Un estimateur de I_a est

$$(8) \quad \frac{gdr}{2\hat{\pi}_a(1 - \hat{\pi}_a)}$$

où gdr représente le taux de différence brut défini par

$$(9) \quad gdr_a = 2 \sum_{a' \neq a} \hat{\pi}_{a,a'}$$

et où $\hat{\pi}_a$ et $\hat{\pi}_{a,a'}$ désignent des estimations de structure latente de π_a et de $\pi_{a,a'}$, respectivement.

Le Bureau of the Census des Etats-Unis (1985, 88-91) a fourni les formules des erreurs-types de même qu'une mesure globale de l'incohérence pour toutes les catégories K combinées, appelée l'indice global d'incohérence, I_{AG} . L'indice global est une mesure de la non-fiabilité au niveau des questions égale à $1 - \kappa$ (Hess, Singer et Bushery 2000) où κ est la mesure de la fiabilité kappa de Cohen (1960) et une moyenne pondérée des indices au niveau des catégories.

Enfin, pour une estimation de π^X donnée, nous pouvons estimer le vecteur K des biais de mesure, notés β_A , associés aux catégories K de A à l'aide de l'identité

$$(10) \quad \beta_A = \pi^{A^A} - \pi^X$$

3.2 Evaluation de la validité de la méthode MLCA

Le présent exposé a comme principal objectif d'évaluer la validité de la stratégie MLCA. Les études précédentes de la mesure de l'erreur de classification de la CPS n'ont pas complètement abordé la validité des stratégies d'estimation utilisées (Meyers 1988). Nous comptons déterminer si la stratégie MLCA est informative et utile pour l'étude de l'erreur de classification de la CPS. Plus précisément, nous espérons déterminer si les estimations modélisées des probabilités d'erreur, $\pi^{A_{1X}}$, reflètent les véritables niveaux d'activité que l'on trouve dans la CPS. Malheureusement, pour des raisons mentionnées antérieurement, il n'existe aucun étalon généralement accepté pour l'évaluation de l'exactitude de la CPS (voir par exemple Sinclair et Caswirth 1996, 1998, Biemer et Forsman 1992, ainsi que Schreiner 1980). Par conséquent, il n'est pas possible d'estimer le biais des estimations MLCA.

Dans la suite du texte, nous examinons la validité des estimations MLCA de l'erreur de classification de la CPS à l'aide de cinq critères:

Nous considérons maintenant les classifications observées de la situation vis-à-vis de l'activité relativement à la CPS notées A_{gi}^* , B_{gi}^* , et C_{gi}^* pour les périodes 1, 2 et 3, respectivement, ou

$$A_{gi}^* = \begin{cases} 1 & \text{si la personne } (g, i) \text{ est classée comme employée au cours de la période } 1 \\ 2 & \text{si la personne } (g, i) \text{ est classée comme sans emploi au cours de la période } 1 \\ 3 & \text{si la personne } (g, i) \text{ est classée comme inactive au cours de la période } 1 \end{cases}$$

avec des définitions semblables pour les indicateurs de réponse B_{gi}^* , et C_{gi}^* pour les périodes 2 et 3, respectivement. À l'aide d'une extension de la notation établie ci-dessus, nous désignons les probabilités de réponse pour chacune de ces classifications par $\pi_{a|g,x} = \Pr(A = a | G = g, X = x)$, avec des définitions semblables pour $\pi_{b|g,y}$ et $\pi_{c|g,z}$. Ainsi, $\pi_{a=1|g,x=2}$ est la probabilité qu'une personne dans le groupe g soit classée par la CPS comme employée ($A = 1$) lorsqu'elle est en réalité sans emploi ($X = 2$). De même, $\pi_{a=2|g,x=2}$ est la probabilité qu'une personne dans le groupe g soit classée correctement comme sans emploi.

Enfin, nous supposons

$$(3) \quad \pi_{a,b,c|g,x,y,z} = \pi_{a|g,x} \pi_{b|g,y} \pi_{c|g,z}$$

ou que l'erreur de classification de la situation observée vis-à-vis de l'activité est indépendante de l'un à l'autre des trois mois. Cette hypothèse, appelée l'hypothèse de l'indépendance locale, a été examinée relativement à la CPS par Meyers (1988) dans son étude de la stratégie d'estimation utilisée par Abowd et Zellmer (1985). Meyers a conclu que l'hypothèse semble représenter une approximation raisonnable. Singh et Rao (1995), qui ont étudié la robustesse de l'hypothèse en fonction d'un certain nombre de scénarios de la population active, ont tiré une conclusion semblable. van de Pol et Langeheine (1997) ont modélisé la distribution combinée des données recueillies au moyen d'un panel et des données de type réinterview à l'aide de modèles de structure latente afin de vérifier l'indépendance locale de divers types de changements de situation vis-à-vis de l'activité. Ils ont trouvé certaines indications que les personnes qui changent de situation vis-à-vis de l'activité manifestent une fiabilité inférieure à celle des personnes qui ne changent pas de situation, l'effet étant toutefois assez faible. Par conséquent, nous allons également supposer (3) sans tenir d'en approfondir davantage la validité dans le présent exposé.

Les classifications de la CPS pour ce qui est de la situation vis-à-vis de l'activité au cours de chaque mois du premier trimestre de l'année sont les variables résultantes de notre analyse. Soit A , B et C , les classifications observées, et X , Y et Z , les véritables classifications (non observées) pour janvier, février et mars, respectivement. Soit G , une variable de groupe (ou stratification) quelconque à définir ultérieurement au cours de l'analyse. Compte tenu de ces hypothèses, nous pouvons formuler la probabilité

3. APPLICATION À LA CPS

3.1 Notation

Une partie de notre évaluation de la stratégie MLCA est une comparaison des estimations MLCA de l'erreur de classification aux estimations tirées de l'analyse des données d'interview-réinterview. Nous utilisons la notation décrite à la section précédente; soit A et A' , la classification de la situation vis-à-vis de l'activité pour l'original et pour la réinterview, respectivement; nous définissons $\pi_a = \Pr(A = a)$ et $\pi_{a'} = \Pr(A' = a')$. Soit AA' , le tableau de tri croisé des interviews-réinterviews observés $K \times K$; soit $\pi_{AA'|X}$, la matrice $K \times K$ des probabilités de cellule, $\Pr(A = a, A' = a' | X = x)$. Si nous supposons que $\pi_{aa'} = \Pr(A = a, A' = a') | X = x) = \pi_{a|x}$, $\pi_{a'} = \Pr(A' = a' | X = x) = \pi_{a'|x}$, nous pouvons poser des mesures parallèles (Bohmstedt 1983), nous

avons

$$(4) \quad \pi_{g,a,b,c} = \sum_{x,y,z} \pi_g \pi_x | g \pi_a | g, x \pi_y | x, g \pi_b | y, g \pi_c | g, z$$

Dans le cadre d'un échantillonnage multinomial, la fonction de vraisemblance pour le tableau $GABC$ est

$$(5) \quad \Pr(GABC) = k \prod_{g,a,b,c} \pi_{gabc}^{g,a,b,c}$$

Les extensions à plus d'une variable de groupe se font directement.

Un membre de l'échantillon de la CPS soit classé dans la cellule (g, a, b, c) du tableau $GABC$ comme suit:

ou $(\pi_{A|X})^T$ désigne la transposée d'un vecteur de probabilités conditionnelles, $\pi_{A|X}$.

Soit π_X , le vecteur K des véritables probabilités de classification. Nous avons alors

$$(6) \quad \pi_{AA'|X} = \pi_{A|X} (\pi_{A|X})^T$$

$$(7) \quad \pi_{AA'} = \pi_{AA'|X} \pi_X$$

c'est-à-dire que la probabilité du tableau de classification produit de la matrice des probabilités de réponses conditionnelles, $\pi_{AA'|X}$, et du vecteur des véritables probabilités de classification, π_X .

et il permet d'ajuster une grande classe de modèles linéaires latentes. Grâce à sa souplesse et à sa généralité, ce logiciel permet à l'analyste de vérifier une gamme considérable de modèles d'erreur de classification et d'examiner les hypothèses au sujet de la cause et des corrélations de l'erreur de classification.

À la section suivante, nous décrivons le modèle MLCA ainsi que la méthode d'estimation et ses fondements théoriques. À la section 3, il est question de la méthode MLCA appliquée à la CPS, et nous ajustons une série de modèles à la CPS tout en examinant l'ajustement de ces modèles. Dans la présente section, nous préparons également des estimations de l'erreur de classification en fonction du meilleur modèle MLCA. À la section 4, nous vérifions de différentes façons la validité des estimations MLCA, notamment en comparant celles-ci aux estimations d'une nouvelle analyse interview-réinterview. Enfin, à la section 5, nous résumons nos résultats et formons des recommandations au sujet de l'utilité de la méthode MLCA pour de futures évaluations de l'erreur de classification de la situation vis-à-vis de l'activité.

2. ANALYSE MARKOVIENNE DE STRUCTURE LATENTE POUR TROIS PÉRIODES DE TEMPS

Les modèles markoviens de structure latente ont d'abord été proposés par Wiggins (1973) et perfectionnés par Poulсен (1982). Van de Pol et de Leeuw (1986) ont établi les conditions dans lesquelles le modèle est identifiable, et ils ont décrit d'autres conditions de l'estimabilité des paramètres de modèle. Dans la présente section, nous élaborons le modèle MLCA dans le contexte de la CPS, et nous en suggérons d'autres applications de même que la généralisation.

Nous divisons la population cible de la CPS en groupes L (âge, race, sexe, par exemple). Soit G_i une variable qui désigne la composition du groupe. Ainsi, $G_i = 1$ si le i -ième membre de la population se trouve dans le groupe 1, $G_i = 2$ pour le groupe 2 et ainsi de suite. Soit X_{gi}^1, X_{gi}^2 et Z_{gi}^3 les véritables classifications de la situation vis-à-vis de l'activité pour la i -ième personne du groupe $G = g$ (pour $g = 1, \dots, L$ et $i = 1, \dots, n^g$) où X_{gi}^g se définit comme

$$X_{gi}^g = \begin{cases} 1 & \text{si la personne } (g, i) \text{ est employée} \\ & \text{au cours de la période 1} \\ 2 & \text{si la personne } (g, i) \text{ est sans emploi} \\ & \text{au cours de la période 1} \\ 3 & \text{si la personne } (g, i) \text{ est inactive} \\ & \text{au cours de la période 1} \end{cases}$$

au cours de la période 1

avec des définitions semblables pour X_{gi}^1 et Z_{gi}^2 au cours des périodes 2 et 3 respectivement. Soit $\pi_{x,y,z|g}$, $\Pr(X=x, Y=y, Z=z | G=g)$

c'est-à-dire qu'au cours de la période 3, la véritable situation d'une personne ne dépend pas de la situation au cours de la période 1, une fois la situation au cours de la période 2 connue. Une autre interprétation est que la situation courante, compte tenu de la situation au cours de la période précédente, ne dépend pas du changement de la période précédente.

On peut concevoir un certain nombre de scénarios dans lesquels l'hypothèse markovienne n'est peut-être pas valable pour la situation d'activité mensuelle. L'hypothèse serait enfreinte, par exemple, si des personnes sans emploi au cours de la période 2 avaient plus de chances d'être sans emploi au cours de la période 3, étant donné qu'elles étaient également sans emploi au cours de la période 1. Le groupe de personnes sans emploi au cours de la période 2 et de la période 1 englobe probablement une plus forte proportion de chômeurs chroniques que le groupe de personnes sans emploi au cours de la période 2 mais non au cours de la période 1. Ce groupe (sans emploi au cours de la période 2 mais non pas de la période 1) compte probablement une plus forte proportion de chômeurs occasionnels en train de changer d'emploi.

Toutefois, la validité de cette hypothèse ne peut pas être examinée convenablement à l'aide des données observées parce que les données comportent une distorsion dont l'ampleur est inconnue à cause de la présence d'erreurs de classification. Il existe au moins deux méthodes d'évaluation de la validité de l'hypothèse markovienne pour des données recueillies au moyen d'un panel. Van de Pol et de Leeuw (1986) ont suggéré une méthode fondée sur quatre cycles de données recueillies au moyen d'un panel qui substitue une restriction markovienne d'ordre 2 à la restriction d'ordre un en (2). Une autre méthode, suggérée par van de Pol et Langheine (1997), fait appel à une combinaison de données de type panel sur la situation d'activité et de données de réinterview pour chaque période de temps. Ni l'une ni l'autre de ces méthodes n'a été utilisée dans le présent exposé pour vérifier directement l'hypothèse de la MLCA. Nous avons plutôt évalué la validité globale des estimations MLCA à l'aide des méthodes décrites à la section 3.2 ci-dessous. À la section 3.6, nous présentons des résultats d'une étude de simulation afin d'illustrer la robustesse des estimations MLCA de l'erreur de classification vis-à-vis des infractions à l'hypothèse markovienne.

(1)
$$\pi_{x,y,z|g} = \pi_{x|g} \pi_{y|g,x} \pi_{z|g,x,y}$$

(2)
$$\pi_{z|g,y} = \pi_{z|g,y}^1$$

$X = y, Z = z | G = g, \pi_{y|g,x}, \pi_{z|g,x,y}, \Pr(Y = y | X = x, G = g)$ et $\pi_{z|g,y}, \Pr(Z = z | Y = y, X = x, G = g)$. La probabilité qu'un membre du groupe g se trouve dans une situation d'activité x au cours de la période 1, y au cours de la période 2 et z au cours de la période 3 est

La méthode de l'analyse markovienne de structure latente, une stratégie prometteuse d'estimation de l'erreur de classification des données d'enquête par panel, n'avait pas encore été appliquée à la CPS. Cette méthode tire avantage de la nature répétitive des enquêtes par panel afin d'extraire de l'information sur l'erreur de classification directe des données d'interview. Le modèle MLCA est en réalité une combinaison de deux modèles: un modèle de chaîne markovienne latente qui représente les changements d'un mois à l'autre de la véritable situation vis-à-vis de l'activité, et un modèle d'erreur de classification qui représente les écarts par rapport à la situation véritable et observée vis-à-vis de l'activité.

Puisque la stratégie MLCA tire avantage de la nature répétitive des enquêtes par panel en vue de l'extraction d'informations sur l'erreur de classification directement des données d'interview, elle n'exige aucune mesure infaillible externe ni aucune mesure obtenue à l'aide de méthodes de réinterview. À cet égard, la méthode offre certains avantages par rapport aux méthodes de Chua et Fuller, Abowd et Zellner, Poterba et Summers, et Sinclair et Gastwirth pour ce qui est de l'évaluation de la qualité des données d'enquête. Dans de nombreuses enquêtes par panel, les réinterviews ne sont pas possibles à cause de restrictions budgétaires, de la complexité des travaux sur place et du fardeau de réponse. La stratégie MLCA est peut-être la seule façon d'évaluer l'erreur de mesure de ces enquêtes. Pour ce qui est des enquêtes par panel comme la CPS, pour lesquelles des données de réinterview sont disponibles, les méthodes de réinterview et MLCA offrent une autre stratégie analytique d'évaluation de l'erreur de classification. Ainsi, comme dans la présente analyse, la méthode MLCA peut servir à modéliser et à vérifier les hypothèses traditionnelles d'analyse des réinterviews. De plus, la stratégie MLCA fournit un schéma statistique pour la combinaison des données recueillies au moyen d'un panel et des données de réinterview en vue de l'obtention d'une information encore plus riche au sujet de l'erreur de classification (van de Pol et Langeheine 1997).

Un autre avantage de la stratégie MLCA est la possibilité d'intégrer l'ensemble des données recueillies au moyen d'un panel aux estimations de l'erreur de classification au lieu de se limiter à l'échantillon relativement petit retenu pour la réinterview. Par conséquent, certains aspects de la qualité des données des enquêtes par panel qu'il n'était pas possible d'examiner jusqu'à présent à cause d'un manque de données sont peut-être abordables maintenant. Le présent exposé décrit nos résultats pour ce qui est de l'utilité de la stratégie de modélisation MLCA pour l'évaluation de l'erreur de classification de la population active dans le cadre de la CPS. Les logiciels servant à ajuster tout un choix de modèles de structure latente de type MLCA ou autre sont accessibles de différentes sources. Le logiciel utilisé dans notre analyse est IEM (Vermunt 1997).

laquelle les réinterviews rapprochées fournissent des valeurs véridiques. Ces auteurs présentent d'importantes indications que les données de réinterview se prêtent à des erreurs de classification appréciables. En réalité, cette prise de conscience a été responsable de l'élimination par le Bureau of the Census du volet réinterview rapprochée du programme de réinterview de la CPS en 1994.

Comme substitut de l'hypothèse d'infaillibilité, Chua et Fuller (1987) et Fuller et Chua (1985) ont appliqué un type de modèle de structure latente aux données de réinterview rapprochées de la CPS afin d'estimer les probabilités de réponse à la CPS. Par souci d'identification du modèle, ces auteurs imposent de rigoureuses restrictions aux probabilités de réponse, forçant le biais dû à l'erreur de classification à être nul tant pour l'interview que pour la réinterview. De plus, ils supposent des erreurs de classification indépendantes pour l'interview et la réinterview (on parle d'hypothèse ICB dans la documentation) et d'un mois à l'autre dans l'échantillon. L'hypothèse ICB représente une limite de leur analyse puisque certains documents indiquent que l'hypothèse n'est peut-être pas valable pour la CPS (voir par exemple O'Muircheartaigh 1991, Singh et Rao 1995). Par conséquent, il se peut que les probabilités de réponse estimées à l'aide de la stratégie de Chua et Fuller soient biaisées.

Sinclair et Gastwirth (1996) et Sinclair et Gastwirth (1998) ont appliqué une stratégie de modélisation de classe latente aux données d'interview-réinterview de la CPS en fonction de restrictions de modèle proposées originellement par Hui et Walter (1980) pour des épreuves diagnostiques médicales. À l'aide de données d'interview-réinterview recoupées selon le sexe, Sinclair et Gastwirth ont supposé que les probabilités d'erreur de classification sont égales pour les hommes et les femmes, tandis que les taux d'activité de la population active sont différents pour ces groupes. Puisque les paramètres de modèle absorbent tous les degrés de liberté disponibles pour l'estimation des paramètres, aucun degré de liberté résiduel n'est disponible pour la vérification d'un manquant du modèle. Par conséquent, leur analyse n'indique pas directement si ces hypothèses au sujet du modèle sont valables pour les données de la CPS. Dans une étude des déterminants du biais dû aux groupes de renouvellement, Shockey (1988) a également appliqué une analyse de structure latente à la CPS. Son analyse indique que le biais dû aux groupes de renouvellement signalé d'abord par Bailar (1975) est peut-être causé par l'erreur de réponse suscitée dans le cadre de l'interview. Shockey n'a pas utilisé de données de réinterview, se fondant plutôt sur des méthodes confirmatoires d'analyse par facteurs pour appuyer ses affirmations. Ses taux d'erreur étaient beaucoup plus élevés que les taux signalés par d'autres auteurs, ce qui indique peut-être un biais dû au modèle. Malheureusement, comme pour Sinclair et Gastwirth, les données de Shockey ne permettent pas une vérification complète des hypothèses du modèle utilisé.

Validité de l'analyse markovienne de structure latente pour l'estimation de l'erreur de classification des données sur la population active

PAUL P. BIEMER et JOHN M. BUSHERY¹

RÉSUMÉ

Les auteurs ont surtout cherché à vérifier la validité des estimations MLCa (analyse markovienne de structure latente) de l'erreur de classification de la population active et à évaluer la possibilité que ce type d'analyse remplace les méthodes traditionnelles d'évaluation de la qualité des données. Les auteurs ont analysé des données d'interview de la Current Population Survey (CPS, enquête sur la population active) pour les trois premiers mois de chacune de trois années (1993, 1995, 1996) et ils ont mené une analyse supplémentaire des données de réinterview non rapprochées de la CPS pour à peu près les mêmes périodes de temps. Les données de réinterview représentent une autre stratégie d'estimation de l'erreur de classification de la CPS qui, si on la compare aux estimations de type MLCa, aide à examiner la validité de la stratégie MLCa. Les auteurs abordent cinq dimensions de la validité de la stratégie MLCa: a) diagnostics modélisés, b) validité de l'ajustement du modèle pour trois années de la CPS, c) accord entre les estimations de type modèle et de type réinterview test-retest des probabilités de réponse, d) accord entre les estimations de type modèle et de type réinterview test-retest de cohérence et e) plausibilité des profils d'erreur de classification. De plus, les auteurs examinent la robustesse des estimations MLCa à l'égard d'infractions à l'hypothèse markovienne. La présente analyse ne fournit aucune raison de mettre en doute la validité de la stratégie MLCa. La méthode a donné de bons résultats pour les cinq épreuves de validité.

MOTS CLÉS: Enquêtes par panel; erreur non due à l'échantillonnage; chômage; qualité des données.

1. INTRODUCTION

La Current Population Survey (CPS, enquête sur la population active) est une enquête sur échantillon menée mensuellement auprès des ménages par le Bureau of the Census des États-Unis en vue de la préparation d'estimations de l'emploi, du chômage et d'autres caractéristiques de la population active des États-Unis. Des estimations nationales tirées de la CPS pour ce qui est de la taille, de la composition et de l'évolution de la composition de la population active sont publiées chaque mois par le Bureau of Labor Statistics in Employment and Earnings des États-Unis. Les estimations de la population active tirées de la CPS sont l'un des indicateurs économiques clés du pays; depuis 1942, l'administration fédérale fait appel aux séries de données de la CPS pour suivre l'évolution mensuelle et annuelle de la population active.

Compte tenu de l'importance des séries de données de la CPS pour la politique gouvernementale, l'exactitude des données a été évaluée à plusieurs reprises. Ainsi, depuis le début des années 1950, le Bureau of the Census a dirigé le programme de réinterview de la CPS afin d'évaluer la qualité des données sur la population active. Dans le cadre de ce programme, on tire un petit sous-échantillon (moins de 5%) des répondants de la CPS et on pose de nouveau certaines questions de l'interview originale, notamment des questions sur la population active. Jusqu'en 1994, on a soumis un quart environ de l'échantillon à une réinterview

non rapprochée et les trois quarts restants à une réinterview rapprochée. Le volet réinterview rapprochée, utilisé surtout en 1994, a causé de préoccupations au sujet de la qualité des données. Toutefois, la réinterview non rapprochée se poursuit encore aujourd'hui et sert à l'estimation de la stabilité (ou de la cohérence des réponses). On trouvera dans Forsman et Schreiner (1991) une description détaillée du programme de réinterview de la CPS.

Plusieurs documents préparés par des chercheurs à l'extérieur du Bureau of the Census ont analysé les données du programme de réinterview de la CPS afin d'estimer l'erreur de classification (voir Sinclair et Gastwirth 1996, 1998; Biemer et Forsman 1992; Chua et Fuller 1987; Poterba et Sumners 1986; Abowd et Zellner 1985). Récemment, Poterba et Sumners (1995) ont utilisé des données du programme de réinterview de la CPS afin d'estimer les taux d'erreur de classification de la CPS et d'évaluer l'effet de l'erreur de classification sur les taux de changement de situation vis-à-vis de l'activité. Comme dans le document de 1986, l'analyse plus récente de ces auteurs se fonde sur l'hypothèse que le processus de rapprochement des réinterviews de la CPS fournit des données que l'on peut considérer comme véridiques. Abowd et Zellner (1985) ont adopté une stratégie semblable.

Plusieurs auteurs (par exemple Sinclair et Gastwirth 1996, 1998; Biemer et Forsman 1992; Forsman et Schreiner 1991; Schreiner 1980) ont mis en doute l'hypothèse selon

Afin d'obtenir (3), nous notons que

$$\sum_{i=1}^I \varepsilon_i^2 w_i = N p_z^2 \sigma_z^2 \sigma_w^2 + N \bar{W} \sigma_z^2$$

$$= N p_z^2 \sigma_z^2 \sigma_w^2 + (1 - p_{y,p}^2) \sigma_y^2 N \bar{W}$$

et

$$\sum_{i=1}^I \varepsilon_i w_i = N p_z \sigma_z^2 \sigma_w^2$$

BIBLIOGRAPHIE

COCHRAN, W. G. (1977). *Sampling Techniques*, 3^e éd. New York: Wiley.

GABLER, S., HAEDER, S., et LAHIRI, P. (1999). Justification à base de modèle de la formule de Kish pour les effets de plan de sondage liés à la pondération et à l'effet de grappe. *Techniques d'enquête*, 25, 1, 119-120.

KISH, L. (1965). *Survey Sampling*. New York: Wiley.

KISH, L. (1992). Weighting for unequal P_i . *Journal of Official Statistics*, 8, 2, 183-200.

SÄRNDAAL, C.-E., SWENSSON, B., et WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Notons, à titre de cas spécial, que si nous fixons $\beta_{y,p} = 0$, qui est le cas de la « pondération au hasard » (Kish 1992), l'estimation de l'effet de plan de sondage se réduit à

$$(7) \quad 1 + rvw(\alpha^2/\delta^2y).$$

Cette estimation se rapproche de celle de Kish pour $\alpha/\delta y$ presque nul.

4. ÉCHANTILLONNAGE SANS REMISE

Le calcul de l'effet de plan de sondage exact pour un échantillonnage sans remise serait plus complexe, car il supposerait que l'on tienne compte des probabilités de sélection composées pour des paires d'unités. Un élargissement heuristique des résultats, par contre, est facile. Rappelons que le rapport entre la variance d'une moyenne simple et la variance pour un échantillonnage aléatoire simple sans remise est approximativement $(1 - n/N)$. Les résultats que nous avons calculés pour l'effet de plan de sondage s'appliquent à des échantillons à un degré à probabilités inégales de n unités sans remise si la variance de l'estimateur de Horvitz-Thompson du total est approximativement $(1 - n/N)$ fois la variance en (2), P'_i étant considéré comme n^{-1} fois la probabilité de sélection globale pour l'unité i (Särndal, Swensson et Wretman 1992, 154).

5. CALCUL DE LA FORMULE DE LA VARIANCE (3)

À partir de (2) nous avons $V(Y) = n^{-1}(\sum_{i=1}^N y'_i/P'_i - Y^2)$. Ensuite, nous notons que (1) suppose que

$$(8) \quad Y^2 = (N\alpha + \beta)^2 = N^2\alpha^2 + 2N\alpha\beta + \beta^2$$

et que

$$\sum_{i=1}^N y'_i/P'_i = \sum_{i=1}^N [\alpha^2/P'_i + \beta^2 P'_i + \alpha\beta/P'_i + 2\alpha\beta + 2\alpha\epsilon_i/P'_i + 2\beta\epsilon_i]$$

$$\alpha^2 = \sum_{i=1}^N P'_i + \beta^2 + \sum_{i=1}^N \epsilon_i^2/P'_i + 2N\alpha\beta + 2\alpha \sum_{i=1}^N \epsilon_i/P'_i$$

$$\alpha^2 n = \sum_{i=1}^N w'_i + \beta^2 + n \sum_{i=1}^N \epsilon_i^2 w'_i + 2N\alpha\beta + 2\alpha n \sum_{i=1}^N \epsilon_i w'_i. \quad (9)$$

Le fait de soustraire (8) de (9) et de diviser par n donne

$$V(Y) = \alpha^2 \left(\sum_{i=1}^N w'_i - N^2/n \right) + \sum_{i=1}^N \epsilon_i^2 w'_i + 2\alpha \sum_{i=1}^N \epsilon_i w'_i.$$

À l'aide de la formulation de régression (1), nous pouvons réexprimer la variance comme suit

$$(2) \quad V(Y) = n^{-1} \sum_{i=1}^N P'_i (y'_i/P'_i - Y^2).$$

$$V(Y) = \alpha^2 N(\bar{w} - N/n) + (1 - \rho_{y,p}^2) \sigma_y^2 N \bar{w} + N \rho_{y,p}^2 \sigma_{\epsilon}^2 w$$

$$(3) \quad + 2\alpha N \rho_{\epsilon w} \sigma_{\epsilon}^2 \sigma_w,$$

Cette expression ne se fonde sur aucune hypothèse au sujet de l'ajustement du modèle de régression (voir le calcul à la section 5).

Si le modèle de régression est assez bien ajusté pour que $\rho_{\epsilon w}$ et $\rho_{\epsilon y}$ soient nuls, la variance en (3) est simplifiée de façon à donner $V(Y) = \alpha^2 N(\bar{w} - N/n) + (1 - \rho_{y,p}^2) \sigma_y^2 N \bar{w}$.

Si l'on avait eu recours à un échantillonnage aléatoire simple avec remise, la variance aurait été $n^{-1} N^2 \sigma_y^2$. Par conséquent, si $\rho_{\epsilon w}$ et $\rho_{\epsilon y}$ sont négligeables, l'effet de plan de sondage est approximativement:

$$(4) \quad \text{deff} = (1 - \rho_{y,p}^2) n \bar{w} / N + (\alpha / \sigma_y)^2 (n \bar{w} / N - 1).$$

Cette approximation n'exige pas que les résidus de la régression soient négligeables, et elle reste valable lorsque σ_{ϵ} est grand. Un examinateur a fait remarquer que la condition voulant que $\rho_{\epsilon w}$ et $\rho_{\epsilon y}$ soient négligeables peut sembler anormale dans un modèle qui régresse y sur P plutôt que sur $w \propto 1/P$. À noter, toutefois, qu'en présence non seulement d'une corrélation nulle entre ϵ et P mais également d'indépendance, nous aurions une corrélation nulle entre les fonctions de ϵ et les fonctions de P , de sorte que $\rho_{\epsilon y}$ et $\rho_{\epsilon w}$ seraient nuls eux aussi.

3. ESTIMATION DE L'EFFET DE PLAN DE SONDAGE

Àfin d'estimer l'effet de plan de sondage lorsque l'échantillon est disponible, nous pouvons utiliser $1 + rvw$ pour estimer $n \bar{w} / N$. Afin d'en comprendre le bien-fondé, notons d'abord que

$$(5) \quad 1 + rvw = \frac{n^{-1} \sum_{i=1}^N w'_i}{\bar{w}^2}.$$

L'espérance du numérateur est $N \bar{w} / n$. L'espérance de $n \bar{w} / N$, et donc le dénominateur de (5) peut être considéré comme un estimateur de $(N/n)^2$. Le fait de diviser l'espérance du numérateur par $(N/n)^2$ nous permet d'obtenir $n \bar{w} / N$. Ainsi, l'effet de plan de sondage peut être estimé à partir de l'échantillon comme suit:

$$(6) \quad (1 - \rho_{y,p}^2)(1 + rvw) + (\alpha/\delta y)^2 (rvw).$$

Un effet de plan de sondage approximatif pour une pondération inégale en cas de corrélation possible entre les mesures et les probabilités de sélection

BRUCE D. SPENCER¹

RÉSUMÉ

On estime couramment l'effet de plan de sondage attribuable à la pondération par 1 plus la variance relative des poids de l'échantillon. Cette formule a été justifiée en l'absence de corrélation entre les probabilités de sélection et la variable d'intérêt. Une approximation de l'effet de plan de sondage est présentée pour tenir compte de la présence d'une corrélation.

MOTS CLÉS : Pondération; effet de plan de sondage; variance d'échantillonnage; échantillons complexes.

1. INTRODUCTION

Une pratique courante consiste à pondérer les observations dans un échantillon à probabilités inégales par les valeurs inverses des probabilités de sélection. On le fait parce que l'absence des poids entraîne une erreur systématique s'il y a une corrélation entre les poids d'échantillonnage et la variable d'intérêt. Un inconvénient du processus de pondération est l'augmentation de la variance d'échantillonnage si les poids varient de façon excessive dans l'échantillon. Cette augmentation peut être quantifiée à l'aide de l'effet de plan de sondage. Celui-ci est le rapport entre la variance de la statistique d'intérêt pour le plan de sondage en question et la variance de la statistique pour un échantillonnage aléatoire simple et la même taille d'échantillon (Kish 1965). Les effets de plan de sondage sont importants tant pour l'approximation des erreurs types une fois que l'échantillon est disponible que pour la prédiction préalable des erreurs types, étape critique d'un plan de sondage efficace.

Kish (1965, 1992) a décrit une approximation de l'effet de plan de sondage pour des estimations pondérées tirées d'échantillons à probabilités inégales: $1 + rvw$, où rvw est défini comme la variance relative des poids de l'échantillon. Si donc w_i est le poids de l'unité i de l'échantillon, et si $rw = n^{-1} \sum_{i=1}^n (w_i - \bar{w})^2 / \bar{w}^2$, est la moyenne de l'échantillon, $rvw = n^{-1} \sum_{i=1}^n (w_i - \bar{w})^2 / \bar{w}^2$. Gabler, Haeder et Lahiri (1999) ont utilisé un modèle de superpopulation pour obtenir un effet de plan de sondage en présence d'une mise en grappes également. Leur formule, qui correspond à des résultats fondés sur un plan de sondage figurant dans Kish (1965), se réduit à $1 + rvw$ en présence d'une corrélation intraclass nulle. L'approximation $1 + rvw$ pour l'effet de plan de sondage se fonde sur un modèle ou un plan dans lequel il n'y a pas de corrélation entre les poids et la variable d'intérêt (de sorte qu'une estimation non pondérée serait tout aussi bonne ou meilleure que l'estimation pondérée). Nous présentons ici une approximation de l'effet de plan de sondage pour un

modèle dans lequel il peut y avoir une corrélation. Ce faisant, nous ne supposons pas que la population est tirée d'une superpopulation. L'exactitude de l'approximation dépend uniquement des caractéristiques du plan de sondage et de la population d'intérêt.

Par souci de simplicité, nous avons recours à l'échantillonnage à un degré à probabilités inégales avec remise. L'élargissement heuristique des résultats à l'échantillonnage sans remise est abordé à la section 4.

2. REPRÉSENTATION DE RÉGRESSION DU PLAN DE SONDAJE ET DE LA POPULATION

Soit y_i la mesure d'intérêt, P_i la probabilité de sélection (d'un tirage à l'autre) pour un échantillon de taille n , et $w_i = 1/(nP_i)$, le poids d'échantillonnage pour l'unité i d'une population de taille N , $1 \leq i \leq N$. À noter que $\bar{P} = \sum_{i=1}^N P_i / N = N^{-1}$. Considérons la droite de régression des moindres carrés

$$y_i = \alpha + \beta P_i + \varepsilon_i \quad (1)$$

où $\alpha = \bar{Y} - \beta / N$, $\beta = \sum_{i=1}^N (y_i - \bar{Y})(P_i - \bar{P}) / \sum_{i=1}^N (P_i - \bar{P})^2$ et $\bar{Y} = \sum_{i=1}^N y_i / N$. Nous notons les variances de population des y_i , des P_i , des ε_i et des w_i par σ_y^2 , σ_P^2 , σ_ε^2 et σ_w^2 avec, par exemple, $\sigma_y^2 = \sum_{i=1}^N (y_i - \bar{Y})^2 / N$. Nous notons la corrélation entre ε_i et w_i par $\rho_{\varepsilon w}$, entre ε_i et P_i par $\rho_{\varepsilon P}$, et entre ε_i et w_i par $\rho_{\varepsilon w}$. Il s'ensuit des propriétés des moindres carrés, ou de façon équivalente des définitions de α et de β , que $\sum_{i=1}^N \varepsilon_i P_i = 0$ et que $\sigma_\varepsilon^2 = (1 - \rho_{\varepsilon P}^2) \sigma_y^2$. Si des données sont disponibles, nous pouvons ajuster la représentation de régression (1) et estimer α , β , σ_ε^2 , et $\rho_{\varepsilon P}$, par exemple, par $\hat{\alpha}$, $\hat{\beta}$, $\hat{\sigma}_\varepsilon^2$ et $\hat{\rho}_{\varepsilon P}$. Soit $\hat{Y} = \sum_{i=1}^n w_i y_i$, l'estimateur pondéré habituel du total de la population, \hat{Y} . La variance de \hat{Y} est bien connue (Cochran 1977, 253):

¹ Bruce D. Spencer, Department of Statistics and Institute for Policy Research, 2006 Sheridan Road, Northwestern University, Evanston, IL 60208, U.S.A.

REMERCIEMENTS

L'auteur tient à remercier la division des données régionales et administratives de Statistique Canada qui a rendu possible ce travail. Il tient aussi à remercier Eric Rancourt, les deux arbitres et le rédacteur associé pour leurs commentaires utiles qui ont permis d'améliorer la qualité de cet article.

BIBLIOGRAPHIE

- BEAUMONT, J.-F., et DEMNATI, A. (1998). Estimation des paramètres d'un mélange fini de distributions pour des données longitudinales dichotomiques; une comparaison d'algorithmes. *Recueil: Symposium 98, Analyse longitudinale pour les enquêtes complexes, Statistique Canada*, 207-213.
- BEAUMONT, J.-F. (1999). A robust estimation method in the presence of nonignorable nonresponse. *Proceedings of the Section on Survey Research Methods, American Statistical Association*. (À paraître).
- D'AGOSTINO, R.B. (1986). Graphical analysis. *Goodness-of-fit Techniques*, (Ed. R.B. D'Agostino et M.A. Stephens), 7-62. New York: Marcel Dekker.
- DEMSTER, A.P., LAIRD, N.M., et RUBIN, R.B. (1977). Maximum likelihood from incomplete data via the EM-algorithm. *Journal of the Royal Statistical Society B*, 39, 1-38.

- GAGNON, F., LEE, H., RANCOURT, E., et SÄRNDAAL, C.-E. (1996). Estimating the variance of the generalized regression estimator in the presence of imputation for the Generalized Estimation System. *Recueil 1996 de la section des méthodes d'enquête, Société statistique du Canada*, 151-156.
- GREENLEES, J.S., REECE, W.S., et ZIESCHANG, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, 77, 251-261.
- LITTLE, R.J.A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.
- RANCOURT, E., LEE, H., et SÄRNDAAL, C.-E. (1994). Corrections du biais pour des estimations d'enquête tirées de données comprenant des valeurs imputées par quotient par suite d'une non-réponse selon un mécanisme confondu. *Techniques d'enquête*, 20, 143-153.
- SÄRNDAAL, C.-E., SWENSSON, B., et WRETMAN, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- SAS INSTITUTE INC. (1990). *SAS/STAT User's Guide*, 2. Version 6, 4^e édition, Cary, NC: SAS Institute Inc.
- ZEGER, S.L., LIANG, K., et ALBERT, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 44, 1049-1060.

6. DISCUSSION

Lorsque l'hypothèse d'un mécanisme de réponse non-ignorable est réaliste et que l'hypothèse de normalité des erreurs du modèle de régression linéaire (3.2) est justifiée, l'utilisation de la méthode du maximum de vraisemblance peut être appropriée. Par contre, lorsque cette dernière hypothèse n'est pas justifiée, les résultats de l'étude de simulation décrite à la cinquième section montrent que la méthode d'estimation robuste présentée dans cet article lui est préférable.

De plus, Beaumont (1999) présente les résultats d'une étude de simulation qui montrent que la méthode d'estimation proposée dans cet article est robuste autant par rapport à l'hypothèse de normalité des erreurs que par rapport au modèle (3.2). Quant à la méthode du maximum de vraisemblance, elle s'est avérée encore plus sensible à la validité du modèle (3.2) qu'à l'hypothèse de normalité des erreurs. Cette dernière méthode ne devrait donc être utilisée que lorsque toutes les hypothèses associées au modèle (3.2) et (3.3) sont raisonnables.

Evidemment, tous les estimateurs sont peu biaisés en présence d'un très faible taux de non-réponse. De même, lorsque le coefficient de corrélation entre X et Y est très élevé, tous les estimateurs sont faiblement biaisés, à l'exception de l'estimateur qui suppose un mécanisme de réponse uniforme $\mu_{P,U}^*$. Dans l'un ou l'autre de ces cas, le choix d'un estimateur devrait être basé sur le critère de simplicité, qui favorise les estimateurs de la section (3.2) et plus particulièrement l'estimateur $\mu_{L,X}^*$.

Il est à noter que les modèles (3.2) et (3.3) pourraient être complexifiés selon la nature du problème. Par exemple, on pourrait inclure d'autres variables indépendantes dans ces modèles. On pourrait également catégoriser la variable X en se servant de variables indicatrices et utiliser ces variables indicatrices dans le modèle (3.3) au lieu de la variable X elle-même.

Dans cet article, on s'est attardé uniquement au problème de l'estimation d'un moyen lorsque le mécanisme de réponse n'est pas ignorable. Les méthodes présentées à la troisième et à la quatrième section sont toutefois applicables à d'autres types d'estimation. Par exemple, on pourrait utiliser les poids ou les valeurs imputées pour l'estimation des paramètres d'une régression quelconque.

L'objectif de cet article a consisté à présenter une méthode d'estimation robuste par rapport à l'hypothèse de normalité des erreurs du modèle (3.2), qui permet de réduire le biais dû à un mécanisme de réponse qui n'est pas ignorable. Dans un travail futur, il serait intéressant d'évaluer des méthodes simples d'estimation de la variance en présence de données imputées et lorsqu'on utilise cette méthode d'estimation robuste.

Le tableau 1 présente les résultats de l'étude de simulation. L'analyse de ce tableau indique que, peu importe la distribution des erreurs, le biais relatif et l'erreur quadratique moyenne de tous les estimateurs est plus faible lorsque la corrélation entre X et Y est plus élevée, ce qui n'est pas surprenant.

Tableau 1

Résultats des simulations permettant de comparer

Estimateur	$R^2 = 80\%$		$R^2 = 50\%$	
	BR (%)	ET	BR (%)	ET
Population avec les erreurs normalement distribuées				
μ_U^*	16,90	0,03	0,84	26,68
$\mu_{P,U}^*$	5,65	0,02	18,02	0,03
$\mu_{P,X}^*$	-0,14	0,03	1,27	0,10
$\mu_{P,X,ML}^*$	1,14	0,03	10,12	0,06
$\mu_{P,X,ROB}^*$	5,50	0,02	17,74	0,03
$\mu_{L,X}^*$	0,13	0,03	0,04	1,03
$\mu_{L,X,ML}^*$	0,64	0,03	0,05	7,53
$\mu_{L,X,ROB}^*$	0,49	0,02	0,04	4,07
μ_U^*	17,83	0,04	0,86	26,60
$\mu_{P,U}^*$	5,44	0,02	16,06	0,04
$\mu_{P,X}^*$	-0,54	0,02	0,04	5,18
$\mu_{P,X,ML}^*$	1,31	0,02	7,43	0,03
$\mu_{P,X,ROB}^*$	5,19	0,02	15,41	0,03
$\mu_{L,X}^*$	-3,42	0,03	0,17	-25,47
$\mu_{L,X,ML}^*$	0,49	0,02	0,04	4,07

L'analyse du biais relatif indique que la méthode du maximum de vraisemblance est la meilleure lorsque les erreurs sont normalement distribuées, suivie par la méthode d'estimation robuste décrite à la section (4.1). Les estimateurs qui supposent un mécanisme de réponse non-ignorable ont un biais relatif plus faible que ceux qui supposent à tort un mécanisme de réponse ignorable. Parmi ces derniers estimateurs, l'estimateur $\mu_{P,U}^*$ est le plus biaisé. Pour une méthode donnée, il existe, en général, peu de différences entre l'estimateur pondéré (2.1) et l'estimateur comprenant des valeurs imputées (2.2). Il faut toutefois donner un léger avantage à ce dernier.

Les conclusions du paragraphe précédent s'appliquent toujours lorsque les erreurs sont exponentiellement distribuées à l'exception que la méthode d'estimation robuste devient la meilleure. Cette constatation n'est pas surprenante puisque la méthode du maximum de vraisemblance est basée sur l'hypothèse de normalité des erreurs. Toutefois, l'estimateur pondéré $\mu_{P,X,ML}^*$ reste faiblement biaisé, ce qui est plus difficilement explicable.

Les conclusions tirées à partir de l'analyse du biais relatif s'appliquent toujours lorsque l'erreur quadratique moyenne. En effet, on remarque que les estimateurs qui sont très biaisés ont une forte tendance à avoir une erreur quadratique moyenne élevée et vice-versa.

Comme à la section (3.3), une fois que les paramètres des modèles (3.2) et (3.3) sont estimés, on peut choisir l'estimateur de la moyenne (2.1) ou (2.2). L'estimateur (2.1) est obtenu en remplaçant w_{ij}^* par $1/p^*(X_j)$, où $p^*(X_j)$ est la probabilité de réponse estimée. Cet estimateur sera noté $\mu_{p^*, Y, ROB}$. L'estimateur (2.2) est également obtenu comme à la section (3.3) en déterminant les valeurs imputées Y_j^* de telle sorte qu'on minimise $\sum_{i \in S} e_i^2$ et que les contraintes $\sum_{i \in S} e_i^* X_i = 0$ et $\sum_{i \in S} e_i^* X_i = 0$ soient respectées, où $e_i^* = Y_i^* - B_0^* - B_1^* X_i$, pour $i \in R$, et $e_i^* = Y_i^* - B_0^* - B_1^* X_i$, pour $i \in O$. Cet estimateur sera noté $\mu_{Y^*, Y, ROB}$. La qualité de ces deux estimateurs de la moyenne dépendra en grande partie de la validité des modèles (3.2) et (3.3) et de la qualité de l'approximation (4.3).

Une modification de l'étape (5) de l'algorithme présentée dans cette section a été proposée dans Beaumont (1999). Les résultats d'une étude de simulation montrent que cette modification donne des résultats légèrement supérieurs à ceux obtenus au moyen de la méthode proposée dans cet article. Cependant, cela ne revient plus à utiliser la méthode du maximum de vraisemblance pour estimer les paramètres du modèle (3.3) étant donné que $f(X_j^* | X_j)$ est connue et on ne peut plus utiliser une procédure de régression logistique à l'étape (5). Il faut toutefois mentionner qu'il n'est pas absolument obligatoire d'avoir recours à la méthode du maximum de vraisemblance pour estimer α_0 et α_1 , même si c'est la méthode qui a été privilégiée dans cet article.

4.2 Vérification de l'hypothèse de normalité des erreurs

Dans le but d'utiliser la méthode du maximum de vraisemblance, on peut être intéressé à vérifier l'hypothèse de normalité des résidus puisque les erreurs ne sont pas observées). Lorsqu'il n'y a pas de non-réponses, une méthode classique (D'Agostino 1986, p. 25, équation 2.11) consiste à faire le graphique de $\Phi^{-1}[F_n^*(e_j^*)]$ en fonction des résidus e_j^* , pour $i \in S$, où $\Phi(\cdot)$ est la fonction de répartition d'une variable aléatoire suivant la loi normale standard et $F_n^*(\cdot)$ est la fonction de répartition empirique. Lorsque les erreurs sont normalement distribuées, les points de ce graphique devraient à peu près s'aligner le long d'une droite de pente 1/σ passant par l'origine.

En présence de non-réponse, on peut utiliser la même technique qu'au paragraphe précédent mais on doit estimer la fonction de répartition empirique au moyen des unités répondantes. Puisque les unités de l'échantillon répondent avec probabilités inégales, la fonction de répartition empirique estimée peut être donnée par (Sæmål, Swensson et Wretman 1992, p. 199):

$$F_n^*(e_j^*) = \frac{\sum_{i: e_i^* \leq e_j^*} 1/d^*(X_i^*)}{\sum_{i \in R} 1/d^*(X_i^*)}.$$

5. ETUDE DE SIMULATION

Il est à noter que, dans cette dernière équation, les probabilités de réponse sont estimées contrairement à la formule de Sæmål, Swensson et Wretman dans laquelle les probabilités de sélection sont connues. Ainsi, on peut vérifier l'hypothèse de normalité des erreurs en faisant le graphique $\Phi^{-1}[F_n^*(e_j^*)]$ en fonction des résidus e_j^* , pour $i \in R$. Cette méthode sera valide à condition que $F_n^*(e_j^*)$ estime correctement $F_n(e_j)$, ce qui est le cas lorsque les probabilités de réponse sont correctement estimées. Lorsque la non-réponse est non-ignorable et qu'on se sert de la méthode d'estimation proposée dans cet article, les probabilités de réponse devraient être bien estimées si les modèles (3.2) et (3.3) sont appropriés de même que l'approximation (4.3).

Afin de comparer les estimateurs de la moyenne présentés aux deux sections précédentes, on a effectué une étude de simulation. On a simulé 4 populations de taille 1 000 selon le modèle (3.2) avec $\beta_0 = 2$ et $\beta_1 = 3$. Les variables aléatoires X_j sont indépendantes entre elles et suivent la loi exponentielle de moyenne 1. Les erreurs e_j sont indépendantes entre elles, ne sont pas corrélées avec les X_j et sont de moyenne nulle et de variance σ^2 . Deux populations ont leurs erreurs suivant la loi normale ($e_j \sim \text{Nor}(0, \sigma^2)$) et les deux autres ont leurs erreurs suivant la loi exponentielle de moyenne σ recentrée à 0 ($e_j \sim \text{Exp}(\sigma) - \sigma$). Pour chacune de ces distributions, une population a l'écart-type σ égal à 1,5 correspondant à un coefficient de corrélation (entre X et Y) au carré de 80% ($R^2 = 80\%$) et l'autre a l'écart-type égal à 3 correspondant à un coefficient de corrélation au carré de 50% ($R^2 = 50\%$). Pour chaque population, on a simulé 1 000 échantillons de répondants selon le modèle (3.3) avec $\alpha_1 = 0,5$. Le paramètre α_0 est déterminé séparément pour chacune des 4 populations de telle sorte que le taux de réponse moyen soit de 70%. Ce paramètre varie entre -1,185 et -0,958. Il est à noter qu'on effectue ici un recensement ($n = N = 1 000$). Ceci à l'avantage de se concentrer seulement sur l'erreur due à la non-réponse puisque il n'y a aucun erreur due à l'échantillonnage. De plus, le fait de générer des populations de taille relativement grande (1 000) permet de mettre surtout l'emphasis sur le biais des estimateurs plutôt que sur leur variance puisque la variance devrait diminuer à mesure que la taille de population augmente (pour un taux de réponse moyen fixe).

Pour chacun des 1 000 échantillons de répondants, on a calculé les 7 estimations de la moyenne décrites aux deux sections précédentes. On a ensuite calculé, pour chaque population, la moyenne et la variance de ces 1 000 estimations, notées $\bar{\mu}$ et $\bar{\sigma}^2$, respectivement. Finalement, on a calculé une estimation du biais relatif (exprimé en pourcentage), $\text{BR}^* = [(\bar{\mu}^* - \mu)/\mu] \times 100\%$, une estimation de l'erreur-type associée à ce biais relatif, $\text{ET}^* = (100/\mu) \sqrt{\bar{\sigma}^2/1 000}$, et une estimation de la racine carrée de l'erreur quadratique moyenne, $\text{REQM}^* = \sqrt{\bar{\sigma}^2 + (\bar{\mu}^* - \mu)^2}$.

de meilleures approximations que (4.3) quoique, dans ce cas, il puisse être préférable d'utiliser la méthode du maximum de vraisemblance.

Une autre propriété intéressante de l'approximation (4.3) est que résoudre alternativement les systèmes d'équations (4.1) et (4.2) peut être obtenu au moyen de l'algorithme suivant:

1. déterminer des valeurs initiales pour les probabilités de réponse (ou pour les paramètres α_0 et α_1 , par exemple) poser $p(Y_i^{(0)}) = 1$ pour toutes les unités répondantes;
2. poser $f = 1$, où f indique le nombre d'itérations;
3. résoudre le système d'équations (4.1) au moyen des estimations courantes des probabilités de réponse, $p(Y_i^{(f-1)})$, en utilisant une procédure de régression pondérée pour ainsi obtenir $\beta_0^{(f)}$ et $\beta_1^{(f)}$;
4. imputer les valeurs manquantes par $Y_i^{(f)} = \beta_0^{(f)} + \beta_1^{(f)} X_i$ pour $i \in O$;
5. résoudre le système d'équations (4.2) en utilisant une procédure de régression logistique pour obtenir $p(Y_i^{(f)})$;
6. arrêter si la convergence est atteinte, sinon poser $f = f + 1$ et retourner à l'étape 3.

Il suffit donc uniquement de disposer d'une procédure de régression linéaire et d'une procédure de régression logistique pour obtenir les estimations désirées. En pratique, cet algorithme est très efficace pour trouver la solution quoique, dans certains cas, il puisse prendre beaucoup d'itérations avant de converger. En fait, il a convergé dans tous les cas où il a été utilisé. Il est également à noter que cet algorithme possède certaines ressemblances avec l'algorithme EM de Dempster, Laird et Rubin (1977), à l'exception qu'ici on ne maximise pas une fonction de vraisemblance.

Dans les simulations de la prochaine section, on a plutôt opté pour l'algorithme Newton-Raphson qui est plus rapide à converger. Cependant, on a dû avoir recours à l'algorithme ci-haut pour les quelques cas pour lequel l'algorithme Newton-Raphson a éprouvé des problèmes de convergence.

L'algorithme proposé pourrait s'avérer très utile pour fournir des valeurs initiales à un algorithme plus rapide tel que l'algorithme Newton-Raphson. Il suffirait d'utiliser l'algorithme proposé avec un critère de convergence peu sévère de telle sorte qu'après quelques itérations seulement, il puisse fournir des valeurs initiales suffisamment bonnes pour assurer la convergence de l'algorithme Newton-Raphson. Dans un autre contexte, Beaumont et Demnati (1998) utilisent une approche similaire qui consiste à commencer le processus itératif avec un algorithme de type EM pour ensuite fournir des valeurs initiales à un algorithme plus rapide de type Newton-Raphson. Ils montrent empiriquement que la combinaison de ces deux algorithmes fournit un bon compromis entre le temps d'exécution et l'efficacité pour trouver la solution.

De même, si la fonction de densité de probabilité $f(X_i, Y_i)$ était connue (pas nécessairement normale mais ne doit pas dépendre des paramètres du modèle 3.3), alors on pourrait estimer les paramètres α_0 et α_1 du modèle (3.3) par la méthode du maximum de vraisemblance, par exemple, et résoudre le système d'équations

$$\sum \frac{\partial \ln p(X_i)}{\partial \alpha_k} = 0 \quad \text{et} \quad \sum \frac{\partial \ln [1 - E(p(X_i) | X_i)]}{\partial \alpha_k} = 0, \quad (4.2)$$

Ainsi, les estimations des paramètres $\beta_0, \beta_1, \alpha_0$ et α_1 sont obtenues en résolvant les équations d'estimation non biaisées (4.1) et (4.2). Un algorithme qui permet de trouver la solution consiste à résoudre alternativement les systèmes d'équation (4.1) et (4.2) jusqu'à ce que la convergence soit atteinte. Pour cela, il faut être en mesure de calculer $E(p(X_i) | X_i)$ dans l'équation (4.2). Cette dernière espérance requiert de connaître la distribution des erreurs ϵ_i , qui est vraisemblablement inconnue. Pour faire face à ce problème, on doit utiliser une approximation et plusieurs peuvent être envisagées, dont l'approximation (3.5). Il serait également possible de développer une stratégie basée sur la méthode du bootstrap en choisissant les unités répondantes proportionnellement à l'inverse de leur probabilité de réponse. Cependant, cette méthode requiert un temps d'exécution important du point de vue informatique et ne sera pas considérée dans ce qui va suivre. Dans cet article, on a plutôt opté pour l'approximation suivante obtenue en linéarisant $p(X_i)$ au moyen d'une série de Taylor évaluée au point $E(Y_i | X_i)$ et en prenant l'espérance des deux premiers termes de cette série:

$$E(p(X_i) | X_i) \approx p(E(Y_i | X_i)) = p(\beta_0 + \beta_1 X_i). \quad (4.3)$$

Il est à noter que l'espérance du deuxième terme de la série est nulle. Cette approximation possède l'avantage de n'exiger que le premier moment de la distribution de Y_i conditionnelle à X_i . En ce sens, elle devrait être robuste par rapport à l'hypothèse de normalité des erreurs puisque elle ne requiert pas de spécifier la distribution des erreurs. Bien entendu, si la distribution des erreurs est connue ou peut être bien estimée, il est possible de trouver

α_0 et α_1 par la méthode du maximum de vraisemblance. On peut aussi utiliser le modèle (3.2). Cependant, les estimations des paramètres ne seront pas convergentes parce que $E(e_j | R_j = 1)$ et $E(e_j X_j | R_j = 1)$ ne sont pas nulles. Même si on disposait d'estimations convergentes, on ne pourrait pas imputer les valeurs manquantes de la façon décrite à la section (3.2) puisque $E(R_j | R_j = 0, X_j) \neq \beta_0 + \beta_1 X_j$ (Greenlees, Reece et Zieschang 1982). Par exemple, la probabilité de réponse est corrélée positivement avec la variable d'intérêt X alors, pour une valeur de X donnée, la moyenne des unités non-répondantes sera inférieure à celle des unités répondantes et sera donc inférieure à la moyenne de toutes les unités réunies. Un raisonnement similaire peut être fait si la probabilité de réponse est corrélée négativement avec la variable d'intérêt. On peut d'ailleurs montrer que

$$E(X_j | R_j = 0, X_j) = \beta_0 + \beta_1 X_j - \frac{\text{cov}(X_j, p(X_j | X_j))}{\text{cov}(X_j, p(X_j | X_j))},$$

$$\text{ou } p(X_j) = P(R_j = 1 | X_j).$$

Les deux approches de la section (3.2) ne sont donc pas valides lorsque le mécanisme de réponse n'est pas igno-

rabile. Dans une telle situation, une meilleure approche consiste à estimer simultanément les paramètres des modèles (3.2) et (3.3). La méthode du maximum de vraisemblance peut être utilisée à cette fin. Cette méthode requiert cependant l'hypothèse supplémentaire que les erreurs e_j soient normalement distribuées (ou toute autre distribution pertinente pour le type de données à analyser), aient une variance constante σ^2 et soient indépendantes entre elles. Le logarithme naturel de la fonction de vraisemblance l peut être écrit:

$$l = \sum_{i \in R} \ln [p(X'_i) f(X'_i | X'_i)]$$

$$+ \sum_{i \in O} \ln [1 - E(p(X'_i) | X'_i)], \quad (3.4)$$

où $f(X'_i | X'_i)$ est la fonction de densité de probabilité d'une loi normale de moyenne $\beta_0 + \beta_1 X'_i$ et de variance σ^2 . La méthode du maximum de vraisemblance consiste à trouver les valeurs des paramètres qui maximisent l . Pour effectuer la maximisation, on doit être en mesure d'approximer $E(p(X'_i) | X'_i)$. On peut y arriver en utilisant une méthode d'intégration numérique comme dans Greenlees, Reece et Zieschang (1982). Dans cet article, on a plutôt opté pour l'approximation suivante (Zeger, Liang et Albert 1988):

$$E(p(X'_i) | X'_i) \approx \frac{1}{1 + \exp \{ -[\alpha_0 + \alpha_1 (\beta_0 + \beta_1 X'_i)] \}}, \quad (3.5)$$

où $k = 1/\sqrt{c^2 \alpha_2^2 + 1}$ et $c = 16\sqrt{3}/15\pi$. Cette approximation est basée sur l'hypothèse que les erreurs sont normalement distribuées et ont une variance constante. On a préféré cette approximation à une méthode d'intégration

Une fois que les paramètres des modèles (3.2) et (3.3) sont estimés, on peut choisir l'estimateur de la moyenne w_{R_i} par $1/p^*(X'_i)$, où $p^*(X'_i)$ est la probabilité de réponse (2.2) peut être obtenu en déterminant les valeurs imputées $X'_{i,ML}$. Cet estimateur sera noté $\mu_{p(X'_i), ML}$. L'estimateur X'_i de telle sorte qu'on minimise $\sum_{i \in S} e_i^2$ et que les contraintes $\sum_{i \in S} e_i = 0$ et $\sum_{i \in S} e_i X'_i = 0$ soient respectées, où $e_i = X'_i - \beta_0 - \beta_1 X'_i$, pour $i \in R$, $e_i = X'_i - \beta_0 - \beta_1 X'_i$, pour $i \in O$, et β_0 et β_1 sont les estimations de β_0 et β_1 respectivement. L'estimateur de la moyenne peut alors s'écrire: $\mu_{l(X'_i), ML} = \beta_0 + \beta_1 \sum_{i \in S} X'_i / n$, où n est la taille de l'échantillon S . La logique derrière cette approche est que les deux contraintes précédentes auraient été respectées si on avait observé la variable X pour toutes les unités de l'échantillon et qu'on avait modélisé cette variable à l'aide du modèle (3.2).

4. MÉTHODE D'ESTIMATION PROPOSÉE

Cette section sert à décrire la méthode d'estimation proposée pour un mécanisme de réponse non-ignorable (section 4.1) de même qu'une méthode graphique (section 4.2) permettant de vérifier l'hypothèse de normalité des erreurs du modèle (3.2).

4.1 Méthode d'estimation pour un mécanisme de réponse qui dépend de X

La méthode du maximum de vraisemblance est valide quand les erreurs sont normalement distribuées et ont la même variance. Lorsque cette hypothèse est violée, il est préférable de recourir à une méthode d'estimation plus robuste. Si les probabilités de réponse $p(X'_i)$ étaient connues et supérieures à zéro pour toutes les unités de l'échantillon, alors une méthode d'estimation robuste (autant par rapport à l'hypothèse de normalité des erreurs que par rapport au modèle 3.2) consisterait à minimiser la somme des erreurs au carré pondérées par l'inverse de la probabilité de réponse $p(X'_i)$. Cette minimisation est équivalente à résoudre le système d'équations

$$\sum_{i \in R} \frac{p(X'_i)}{1} (X'_i - \beta_0 - \beta_1 X'_i) Z_{ik} = 0, \quad k = 1, 2, \quad (4.1)$$

où $Z_{i1} = 1$ et $Z_{i2} = X'_i$. Cette approche est considérée robuste par rapport à l'hypothèse de normalité puisque la méthode des moindres carrés n'exige pas de spécifier la distribution des erreurs. Le fait de pondérer par l'inverse de

Une méthode d'estimation en présence de non-réponse non-ignorable

JEAN-FRANÇOIS BEAUMONT¹

RÉSUMÉ

Lorsque le mécanisme de réponse dans une enquête dépend d'une variable d'intérêt mesurée dans cette même enquête et qui n'est observée que pour une partie de l'échantillon seulement, on dit qu'on est en présence de non-réponse non-ignorable. Dans une telle situation, ne pas tenir compte de la non-réponse peut engendrer un biais important dans l'estimation d'une moyenne ou d'un total. Pour contre ce problème, on peut modéliser conjointement le mécanisme de réponse et la variable d'intérêt et effectuer l'estimation par la méthode du maximum de vraisemblance. La critique principale de cette méthode est que l'estimation par la méthode du maximum de vraisemblance est basée sur l'hypothèse de normalité vérifiable de normalité des erreurs pour le modèle impliquant la variable d'intérêt. Dans cet article, on propose une méthode d'estimation robuste par rapport à l'hypothèse de normalité puisqu'elle est construite de telle sorte qu'elle n'exige pas de spécifier la distribution des erreurs. La méthode est évaluée au moyen de simulations de Monte Carlo. On propose également une méthode simple permettant de vérifier la validité de l'hypothèse de normalité des erreurs quand la non-réponse n'est pas ignorable.

MOTS CLÉS: Non-réponse non-ignorable; maximum de vraisemblance; équations d'estimation; imputation par la régression; pondération.

1. INTRODUCTION

Lorsque le mécanisme de réponse dans une enquête dépend d'une variable d'intérêt mesurée dans cette même enquête et qui n'est observée que pour une partie de l'échantillon seulement, on dit qu'on est en présence de non-réponse non-ignorable. Par exemple, lorsqu'on mesure le revenu, il peut être réaliste de supposer que les personnes ayant un faible revenu ont moins tendance à répondre que les personnes ayant un revenu élevé ou vice-versa. Le lecteur est renvoyé à Little (1982) pour une définition formelle du concept de non-réponse non-ignorable. Dans une telle situation, ne pas tenir compte de la non-réponse peut engendrer un biais important dans l'estimation d'une moyenne ou d'un total. Pour contre ce problème, on peut modéliser conjointement le mécanisme de réponse et la variable d'intérêt, effectuer l'estimation par la méthode du maximum de vraisemblance comme, entre autres, dans Greenlees, Reece et Zieschang (1982) et imputer les valeurs manquantes. La critique principale de cette méthode est que l'estimation par la méthode du maximum de vraisemblance est basée sur l'hypothèse de normalité vérifiable de normalité des erreurs pour le modèle impliquant la variable d'intérêt.

Rancourt, Lee et Særdal (1994) présentent des facteurs de correction simples pour réduire le biais engendré par la non-réponse qui n'est pas ignorable sans recourir à l'hypothèse de normalité et sans modèle pour le mécanisme de réponse. Ces facteurs de correction ne sont toutefois disponibles que pour l'imputation par le quotient.

Dans cet article, on propose une méthode d'estimation robuste par rapport à l'hypothèse de normalité puisqu'elle est construite de telle sorte qu'elle n'exige pas de spécifier

unités répondantes:

L'estimateur de la moyenne, $\mu = \sum_{i \in P} Y_i / N$, où N est la taille de la population, peut être obtenu en pondérant les dants est noté O . On suppose qu'on dispose d'au moins une variable observée pour toutes les unités de l'échantillon et corrélée avec Y .

Dans ce qui va suivre, on tente d'estimer la moyenne de la variable Y pour une certaine population P . Pour y arriver, on sélectionne un échantillon S et la variable Y est observée pour une partie de l'échantillon seulement. L'échantillon des répondants est noté R et l'échantillon des non-répondants est noté O . On suppose qu'on dispose d'au moins une variable observée pour toutes les unités de l'échantillon et corrélée avec Y .

2. NOTATION

La distribution des erreurs. On propose également une méthode simple permettant de vérifier la validité de l'hypothèse de normalité des erreurs quand la non-réponse n'est pas ignorable.

À la deuxième section, on définit le problème et on introduit un peu de notation. La troisième section sert à présenter différents estimateurs de la moyenne d'une population sous différentes hypothèses concernant le mécanisme de réponse et la distribution des données. À la quatrième section, on présente la méthode d'estimation proposée quand la non-réponse n'est pas ignorable. La cinquième section sert à décrire les résultats d'une étude de simulation permettant de comparer les différents estimateurs décrits aux deux sections précédentes. Finalement, la dernière section contient une brève discussion.

(2.1)

- Haines, Pollock et Pantula: Estimation de la taille et des chiffres de population
- FULLER, W.A., et BURMEISTER, L.F. (1972). Estimators for samples selected from two overlapping frames. *Proceedings of the Social Statistics Section, American Statistical Association*, 245-249.
- HAINES, D.E. (1997). Estimating Population Parameters Using Multiple Frame and Capture-Recapture Methodology. Thèse de doctorat, North Carolina State University.
- HAINES, D.E., et POLLOCK, K.H. (1998a). Combinaison de bases multiples pour estimer la taille et les chiffres de la population. *Techniques d'enquête*, 24, 81-9188.
- HAINES, D.E., et POLLOCK, K.H. (1988b). Estimating the number of active and successful bald eagle nests: an application of the dual frame method. *Environmental and Ecological Statistics*, 5, 245-256.
- HANSEN, M.H., HURWITZ, W.N., et MADOW, W.G. (1953). *Sample Survey Methods and Theory I*. New York: John Wiley & Sons.
- HARTLEY, H.O. (1962). Multiple frame surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 203-206.
- HARTLEY, H.O. (1974). Multiple frame methodology and selected applications. *Sankhyā*, 36, 3, C, 99-118.
- HORVITZ, D.G., et THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- HUGGINS, R.M. (1989). On the statistical analysis of capture experiments. *Biometrika*, 76, 133-140.
- KOTT, P.S., et VOGEL, F.A. (1995). Multiple-frame business surveys. *Business Survey Methods* (Ed. B.G. Cox). New York: John Wiley & Sons. 185-203.
- LINCOLN, F.C. (1930). Calculating Waterfowl Abundance on the Basis of Banding Returns. U.S. Department of Agriculture, Circular, 118.
- LUND, R.E. (1968). Estimators in multiple frame surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 282-288.
- NEALON, J.P. (1984). Review of the Multiple and Area Frame Estimators, Staff Report 80. U.S. Department of Agriculture, Statistical Reporting Service, Washington, D. C.
- NORRIS, J.L., et POLLOCK, K.H. (1996). Nonparametric MLE under two closed capture-recapture models with heterogeneity. *Biometrics*, 52, 639-649.
- OTTS, D.L., BURNHAM, K.P., WHITE, G.C., et ANDERSON, D.R. (1978). Statistical inference for capture data on closed animal populations. *Wildlife Monographs*, 62, 1-135.
- PETERSEN, C.G.J. (1896). The yearly immigration of young plaice into the Limfjord from the German Sea. *Rep. Danish Biol. Sta.*, 6, 1-48.
- POLLOCK, K.H. (1991). Modeling capture, recapture, and removal statistics for estimation of demographic parameters for fish and wildlife populations: past, present, and future. *Journal of the American Statistical Association*, 86, 225-238.
- POLLOCK, K.H., HINES, J.E., et NICHOLS, J.D. (1984). The use of auxiliary variables in capture-recapture and removal experiments. *Biometrics*, 40, 329-340.
- POLLOCK, K.H., TURNER, S.C., et BROWN, C.A. (1994). Techniques de saisie-ressaisie pour l'estimation de la taille de la population et de totaux de population lorsqu'on ne dispose pas d'une base de sondage complète. *Techniques d'enquête*, 20, 121-128.
- QUENOUILLÉ, M.H. (1956). Notes on bias in estimation. *Biometrika*, 43, 353-360.
- SEKAR, C.C., et DEMING, W.E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44, 101-115.
- WOLTER, K.M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, 81, 338-346.
- WOLTER, K.M. (1990). Capture-recapture estimation in the presence of a known sex ratio. *Biometrics*, 46, 157-162.

Même si la moyenne de l'erreur-type estimée de N_{p_i} est inférieure à l'écart-type empirique, la différence est relativement faible. De même, les probabilités de couverture de l'intervalle de confiance de 95% fondé sur N_{p_i} sont très proches de 0,95. Par contre, les probabilités de couverture de l'intervalle de confiance de 95% fondé sur N_{CH} sont de 0,271 et 0,009 pour $N = 300$ et $N = 1\ 000$, respectivement. Compte tenu de nos simulations, nous recommandons l'utilisation de N_{p_i} avec une grande valeur de p_i . Le choix de p_i est déterminé dans la pratique par les coûts d'échantillonnage de la base areolaire, qui ne sont pas pris en considération dans notre étude.

3.7.2 Estimation du total de la population

Pour les totaux de la population, nous obtenons des résultats qui sont très semblables à ce que nous avons observé pour la taille de la population. En général, le biais relatif et l'erreur-type diminuent à mesure que p_i augmente et que la taille de la population augmente. Nous remarquons également que l'erreur-type estimée relative moyenne est très proche de l'écart-type empirique de l'estimateur normalisé en fonction du total. Cela indique que la formule approximative de l'erreur-type en (7) représente une bonne estimation de l'erreur-type. À noter également que les probabilités de couverture empiriques se trouvent pour la plupart à trois erreurs-types près de 0,95. Autrement dit, la plupart des probabilités de couverture empiriques se trouvent à $(0,95 \pm 3[0,95 \times 0,05/1\ 000]^{1/2}) = (0,929, 0,971)$ près.

4. SOMMAIRE

Dans le présent exposé, nous avons étudié le rendement de l'estimateur de Horvitz-Thompson estimé de la taille de la population et du total de la population en fonction d'échantillons tirés de listes et de bases areolaires. Nous avons présenté des méthodes d'estimation des paramètres du modèle de régression logistique pour les probabilités d'inclusion. Même si de nombreux modèles et autres estimateurs ont été considérés dans Haines (1997), nous présentons des résultats d'étude de simulation pour deux modèles seulement et quelques estimateurs. Nous croyons que les méthodes utilisées dans le présent exposé pourraient être très utiles aux analystes d'enquête parce que le caractère incomplet des listes est une réalité de tous les jours. Nos résultats comptent parmi les premiers à proposer une méthode d'estimation des chiffres de population qui puisse tenir compte du caractère incomplet et modéliser les probabilités d'inclusion en fonction des covariables.

REMERCIEMENTS

Les auteurs tiennent à remercier le rédacteur et un rédacteur adjoint de remarques qui ont permis d'améliorer

le contenu et la présentation de l'exposé. Les auteurs remercient également Christine Bunck, gestionnaire du programme BEST, Biological Resources Division, U.S. Geological Survey, de l'appui financier accordé à la présente recherche dans le cadre d'un bon de travail remis à la North Carolina State University. Les vues exprimées sont celles des auteurs et ne reflètent pas nécessairement les vues du Bureau of the Census.

BIBLIOGRAPHIE

- ALHO, J.M. (1990). Logistic regression in capture-recapture models. *Biometrics*, 46, 623-635.
- BOSECKER, R.R., et FORD, B.L. (1976). Multiple frame estimation with stratified overlap domain. *Proceedings of the Social Statistics Section, American Statistical Association*, 219-224.
- BURNHAM, K.P. (1972). Estimation of Population Size in Multiple Capture Studies when Capture Probabilities Vary Among Animals. Thèse de doctorat, Oregon State University.
- BURNHAM, W.S. (1978). Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika*, 65, 625-633.
- BURNHAM, K.P., et OVERTON, W.S. (1979). Robust estimation of population size when capture probabilities vary among animals. *Ecology*, 60, 927-936.
- CHAO, A. (1988). Estimating animal abundance with capture frequency data. *Journal of Wildlife Management*, 52, 295-300.
- CHAO, A., LEE, S.-M., et JENG, S.-L. (1992). Estimating population size for capture-recapture data when capture probabilities vary by time and individual animal. *Biometrics*, 48, 201-216.
- CHAPMAN, D.G. (1951). Some Properties of the Hypergeometric Distribution with Applications to Zoological Censuses. University of California, University of California Publication in Statistics.
- COCHRAN, R.S. (1965). *Theory and Applications of Multiple Frame Surveys*. Thèse de doctorat, Iowa State University.
- COCHRAN, W.G. (1977). *Sampling Techniques*. 3-ième édition. New York: John Wiley & Sons.
- COWAN, C.D., et MALEC, D. (1986). Capture-recapture models when both sources have clustered observations. *Journal of the American Statistical Association*, 81, 347-353.
- FAULKNER, G.D., et GAROU, A. (1991). Estimating a population total using an area frame. *Journal of the American Statistical Association*, 86, 445-449.
- FECES, R., TORTORA, R.D., et VOGEL, F.A. (1986). Sampling frames for agriculture in the United States. *Journal of Official Statistics*, 2, 279-292.
- FIENBERG, S.E. (1992). Bibliographie sur la modélisation à l'aide de la saisie-résaisie avec application au redressement des chiffres du recensement pour éliminer le sous-dénombrement. *Techniques d'enquête*, 18, 157-169.

Pour le modèle 2, la probabilité d'inclusion est fonction de la covariable. Par conséquent, N_{CH} n'est pas un estimateur approprié pour N . Nous observons que N_{CH} sous-estime appréciablement la vraie taille de la population. Par contre, N_{P_A} fournit une bonne estimation de N . Le biais dans N_{P_A} diminue à mesure que P_A augmente dans le modèle 2. De plus, le biais relatif diminue à mesure que la taille de la population augmente.

Comme on pouvait s'y attendre, l'écart-type de N_{P_A} diminue à mesure que la probabilité d'inclusion de la base aréolaire P_A augmente. Par exemple, dans le modèle 1 où $N = 300$, l'inclusion d'un échantillon de base aréolaire

5% fait baisser l'écart-type relatif de 0,077 à 0,059, une réduction de 23%. Lorsqu'on utilise un échantillon de base aréolaire de 20%, l'écart-type relatif diminue de 0,077 à 0,048, une réduction de 38%. Lorsque $N = 1\ 000$, l'inclusion d'un échantillon de base aréolaire de 5% fait baisser l'écart-type relatif de 0,035 à 0,030, une réduction de 14%. Le fait d'augmenter l'échantillon de base aréolaire à 20% fait baisser l'écart-type relatif de 0,035 à 0,025, une diminution de 29%. De façon générale, les erreurs-types relatives diminuent à mesure que la taille de la population augmente.

Tableau 4

Estimations du total de la population par sous-échantillonnage mises à l'échelle selon X pour le modèle 1

P_{xy}	0	0,5	1,0	$N = 300$			$N = 1\ 000$		
	Moyenne des estimations	Ecart-type des estimations	Moyenne de l'erreur-type estimée	Couverture	Moyenne des estimations	Ecart-type des estimations	Moyenne de l'erreur-type estimée	Couverture	Moyenne des estimations
P_{xy}	1,004	0,077	0,076	0,953	1,013	0,081	0,081	0,951	1,003
	1,003	0,073	0,072	0,951	1,012	0,080	0,072	0,951	1,002
	1,001	0,062	0,061	0,951	1,008	0,070	0,060	0,949	1,001
	1,002	0,042	0,041	0,946	1,001	0,059	0,041	0,942	1,001
0	1,002	0,065	0,064	0,950	1,009	0,072	0,057	0,950	1,009
	1,008	0,062	0,061	0,950	1,013	0,070	0,051	0,946	1,013
	1,007	0,050	0,051	0,950	1,009	0,062	0,045	0,949	1,009
	1,005	0,034	0,034	0,952	1,004	0,038	0,033	0,944	1,004
0,5	1,002	0,035	0,034	0,954	1,009	0,035	0,033	0,942	1,009
	1,002	0,033	0,033	0,950	1,018	0,030	0,025	0,942	1,018
	1,002	0,033	0,033	0,950	1,012	0,028	0,023	0,942	1,012
	1,002	0,033	0,033	0,950	1,009	0,025	0,020	0,942	1,009
1,0	1,001	0,021	0,020	0,952	1,004	0,021	0,017	0,947	1,004
	1,001	0,020	0,020	0,950	1,003	0,020	0,016	0,947	1,003
	1,001	0,017	0,017	0,951	1,001	0,017	0,012	0,947	1,001
	1,001	0,017	0,017	0,954	1,001	0,017	0,012	0,947	1,001

Tableau 5

Estimations du total de la population par sous-échantillonnage mises à l'échelle selon X pour le modèle 2

P_{xy}	0	0,5	1,0	$N = 300$			$N = 1\ 000$		
	Moyenne des estimations	Ecart-type des estimations	Moyenne de l'erreur-type estimée	Couverture	Moyenne des estimations	Ecart-type des estimations	Moyenne de l'erreur-type estimée	Couverture	Moyenne des estimations
P_{xy}	1,010	0,098	0,094	0,935	1,007	0,061	0,061	0,947	1,003
	1,010	0,092	0,089	0,935	1,007	0,061	0,061	0,947	1,003
	1,009	0,078	0,074	0,926	1,005	0,051	0,051	0,947	1,002
	1,006	0,052	0,051	0,931	1,002	0,034	0,032	0,947	1,000
-0,3	1,003	0,049	0,048	0,955	1,002	0,032	0,027	0,947	1,000
	1,003	0,049	0,048	0,955	1,002	0,032	0,027	0,947	1,000
	1,003	0,049	0,048	0,955	1,002	0,032	0,027	0,947	1,000
	1,003	0,049	0,048	0,955	1,002	0,032	0,027	0,947	1,000
0	1,008	0,065	0,064	0,950	1,002	0,034	0,028	0,954	1,000
	1,008	0,065	0,064	0,950	1,002	0,034	0,028	0,954	1,000
	1,008	0,065	0,064	0,950	1,002	0,034	0,028	0,954	1,000
	1,008	0,065	0,064	0,950	1,002	0,034	0,028	0,954	1,000
0,5	1,002	0,035	0,034	0,954	1,002	0,033	0,018	0,965	1,000
	1,002	0,035	0,034	0,954	1,002	0,033	0,018	0,965	1,000
	1,002	0,035	0,034	0,954	1,002	0,033	0,018	0,965	1,000
	1,002	0,035	0,034	0,954	1,002	0,033	0,018	0,965	1,000
1,0	1,001	0,021	0,020	0,952	1,001	0,017	0,011	0,947	1,000
	1,001	0,021	0,020	0,952	1,001	0,017	0,011	0,947	1,000
	1,001	0,021	0,020	0,952	1,001	0,017	0,011	0,947	1,000
	1,001	0,021	0,020	0,952	1,001	0,017	0,011	0,947	1,000

$$\bar{Y}_A = \sum_{i \in \text{"échantillon de la base aéroclaire"}} \frac{Y_i}{N_A}$$

Il s'agit de l'estimateur de Horvitz-Thompson fondé sur l'échantillon de la base aéroclaire seulement. Puisque le dénombrement complet des segments aéroclaires coûte cher, p_A est typiquement petite dans la pratique. Pour une petite p_A , X_A devrait avoir une variance beaucoup plus grande que X_{p_A} puisque l'estimateur X_{p_A} inclut des renseignements des listes en plus de renseignements des échantillons de la base aéroclaire. Les résultats pour X_A ne sont donc pas inclus.

3.5 Variance estimée de l'estimateur

Dans notre étude de simulation, les valeurs de p_A considérées sont très petites. Par contre, la probabilité d'inclusion dans au moins une des listes est grande pour chaque individu. Par conséquent, π_j est proche de π_j en (11) est proche de π_j/π_j . Ainsi, le second membre des équations (10) et (17), comportant $\pi_j'' - \pi_j/\pi_j$, devrait être petit. Nous n'avons pas inclus ce membre dans notre estimation de la variance. Malgré cette omission, nous remarquons que la variance estimée est très proche de la variance empirique que l'estimateur pour les modèles que nous considérons.

3.6 Statistiques sommaires

Pour les estimations de la taille de la population, nous présentons des moyennes établies pour les 4 000 répétitions correspondant aux quatre valeurs de p_{xy} et les 1 000 répétitions de Monte Carlo pour chaque p_{xy} . Pour chaque modèle, nous résumons la moyenne et l'écart-type des estimations, la racine carrée de l'erreur quadratique moyenne relative (REQMR) en pourcentage et les probabilités de couverture empiriques pour un intervalle de confiance de 95%. Ces mesures sont toutes normalisées selon la taille de la population N . Nous présentons des résultats pour les modèles 1 et 2 dans les tableaux 2 et 3, respectivement.

De même, pour les estimations du total de la population, nous présentons des données statistiques sommaires établies sous forme de moyennes pour les 1 000 répétitions correspondantes à chaque combinaison paramétrique. Nous résumons la moyenne et l'écart-type des estimations de même que la moyenne des erreurs-types estimées des estimateurs, les estimations étant mises à l'échelle en fonction du vrai total (X) pour cette répétition. Autrement dit, pour chaque répétition, nous divisons l'estimation par son total de répétition, X . Nous calculons alors la moyenne et les écarts-types de ces estimations normalisées. De même, pour chaque répétition, nous calculons l'erreur-type estimée de l'estimateur du total divisée par le total pour la répétition, et nous calculons ensuite la moyenne de ces valeurs normalisées. Celles-ci sont signalées parce que les totaux changent d'une répétition à l'autre. Enfin, nous signalons les probabilités de couverture des intervalles de confiance de 95% pour le total. Les résultats pour les modèles 1 et 2 sont présentés dans les tableaux 4 et 5 respectivement.

3.7. Conclusions

3.7.1 Estimation de la taille de la population

Dans le modèle 1, les probabilités d'inclusion ne dépendent pas de la covariable. Dans ce cas, l'estimateur de Chapman N^{CH} est très proche de l'estimateur du maximum de vraisemblance (Lincoln-Petersen) et on s'attend donc à ce qu'il donne de meilleurs résultats que N_0 . L'estimateur N_0 perd de son efficacité puisqu'il sert à estimer les paramètres $\alpha_B, \beta_B, \alpha_{B_2}$ et β_{B_2} , dont la valeur est nulle dans ce modèle. L'estimateur $N_{0.05}$ est à peu près aussi efficace que N^{CH} . Le biais pour toutes les estimations est minime. Pour le modèle 1, nous remarquons que la moyenne de l'écart-type estimé est proche de l'écart-type des estimations. Cela indique que l'estimation de l'erreur-type que nous utilisons donne de bons résultats. De même, nous remarquons que les probabilités de couverture empiriques se trouvent toutes à trois erreurs-types près de 0,95. Autrement dit, toutes les probabilités de couverture empiriques se trouvent à $(0,95 \pm 3 [0,95 \times 0,05/4 000]^{1/2}) = (0,94, 0,96)$ près.

Tableau 3

Estimations de la taille de la population pour le modèle 2

$N = 300$	N^{CH}	N_0	$N_{0.05}$	$N_{0.20}$
-----------	----------	-------	------------	------------

Moyenne des estimations divisée par N	0,922	1,006	1,005	1,003
Ecart-type des estimations divisé par N	0,032	0,052	0,049	0,040
Moyenne de l'écart-type estimé de l'estimateur divisée par N	0,028	0,052	0,048	0,040
REQMR en pourcentage	0,007	0,003	0,002	0,002
Couverture	0,271	0,953	0,954	0,951

$N = 1 000$

Moyenne des estimations divisée par N	0,921	1,001	1,001	1,001
Ecart-type des estimations divisé par N	0,018	0,028	0,027	0,022
Moyenne de l'écart-type estimé de l'estimateur divisée par N	0,015	0,027	0,026	0,021
REQMR en pourcentage	0,007	0,0008	0,0007	0,0005
Couverture	0,009	0,949	0,949	0,949

Tableau 2

Estimations de la taille de la population pour le modèle 1

$N = 300$	N^{CH}	N_0	$N_{0.05}$	$N_{0.20}$
-----------	----------	-------	------------	------------

Moyenne des estimations divisée par N	0,999	1,011	1,007	1,004
Ecart-type des estimations divisé par N	0,059	0,077	0,059	0,048
Moyenne de l'écart-type estimé de l'estimateur divisée par N	0,059	0,072	0,059	0,047
REQMR en pourcentage	0,003	0,006	0,004	0,002
Couverture	0,947	0,955	0,957	0,950

$N = 1 000$

Moyenne des estimations divisée par N	1,000	1,003	1,002	1,002
Ecart-type des estimations divisé par N	0,031	0,035	0,030	0,025
Moyenne de l'écart-type estimé de l'estimateur divisée par N	0,032	0,034	0,030	0,025
REQMR en pourcentage	0,001	0,001	0,001	0,001
Couverture	0,954	0,959	0,958	0,956

d'une covariable x_j . Deuxièmement, nous supposons que la covariable peut être corrélée avec la variable de réponse y_j . De plus, nous supposons que x_j et y_j suivent une distribution normale logarithmique avec corrélation ρ_{xy} . La distribution normale logarithmique est utilisée de façon à adapter une distribution asymétrique des covariables. Nous génerons x_j comme e^{u_j} et y_j comme e^{v_j} , où u_j et v_j sont des variables aléatoires normales à deux dimensions avec moyennes nulles, des variances unitaires et une corrélation ρ_{uv} . Il est possible de montrer que $\rho_{uv} = \log[p_{xy} / (e - 1) + 1]$.

Considérons une population de taille N . Supposons qu'il existe deux listes indépendantes, B_1 et B_2 , et une base aréolaire, A . La base aréolaire est supposée complète en ce sens qu'elle couvre la population entière. On tire un échantillon de segments de la base aréolaire et on observe les unités au sein de chaque segment aréolaire. Soit p_{ij} la probabilité d'inclusion d'un élément quelconque dans l'échantillon de la base aréolaire, où p_{ij} est supposée identique pour tous les individus.

La probabilité que le i -ième élément soit inclus dans la j -ième liste est donnée par le modèle de régression logistique (1) pour $i = 1, \dots, N$ et $j = B_1, B_2$. Nous supposons que la probabilité que le i -ième élément soit inclus dans la liste B_1 et B_2 . Nous utilisons la probabilité p_{ij} pour inclure le i -ième élément dans la liste j . Enfin, à l'aide de $p_{ij} = 0,05$, nous identifions les éléments qui appartiennent à la base aréolaire A . Nous répétons le processus pour le cas $p_{ij} = 0,20$. Pour chaque combinaison paramétrique, nous génerons 1 000 répétitions de Monte Carlo.

3.3 Génération des données

Pour chacun des seize modèles ci-dessus, nous génerons d'abord (x_j, y_j) à l'aide de la distribution normale logarithmique à deux dimensions pour $i = 1, \dots, N$. Nous «génerons» (identifications) ensuite les unités qui appartiennent aux listes B_1 et B_2 . Nous utilisons la probabilité p_{ij} pour inclure le i -ième élément dans la liste j . Enfin, à l'aide de $p_{ij} = 0,05$, nous identifions les éléments qui appartiennent à la base aréolaire A . Nous répétons le processus pour le cas $p_{ij} = 0,20$. Pour chaque combinaison paramétrique, nous génerons 1 000 répétitions de Monte Carlo.

Modèle	α_1	β_1	α_2	β_2	$p_{B_1(E)}$	$p_{B_2(E)}$	$1 - (1 - p_{B_1(E)}) - (1 - p_{B_2(E)})$
1	0	0	0	0	0,5	0,5	0,75
2	-0,5478	0,8	-0,5478	0,8	0,6838	0,6838	0,90

Sommaire des paramètres de modèle

Tableau 1

au moins une des deux listes est donnée par $1 - (1 - p^2)$. Cette relation est utilisée dans le modèle 2. Les valeurs particulières de α_j et β_j pour les deux modèles sont résumées dans le tableau 1.

Pour la taille de la population, nous considérons l'estimateur de Chapman, N_{CH} , donné en (8). Cet estimateur suppose que $\beta_{B_1} = \beta_{B_2} = 0$ et n'utilise pas l'information de l'échantillon de la base aréolaire. Nous considérons également les estimateurs de Horvitz-Thompson estimés dont il est question à la section 2.

Pour l'estimation du total de la population d'une variable de réponse, nous considérons le cas dans lequel la réponse est observée pour tous les éléments de la liste. Les éléments d'une base aréolaire sont échantillonnés avec des probabilités $p_{ij} = 0,05$ et $0,20$. Nous ne considérons pas d'estimations du total de la population fondées sur des sous-échantillons de chaque liste. L'estimation du total de la population, $N_{p_{ij}}$, se présente sous la même forme que (16) avec π_j définie en (15). De même, l'estimation de la taille de la population, $N_{p_{ij}}$, se présente sous la même forme que (9).

3.4 Estimateurs

Pour la taille de la population, nous considérons l'estimateur de Chapman, N_{CH} , donné en (8). Cet estimateur suppose que $\beta_{B_1} = \beta_{B_2} = 0$ et n'utilise pas l'information de l'échantillon de la base aréolaire. Nous considérons également les estimateurs de Horvitz-Thompson estimés dont il est question à la section 2.

Pour l'estimation du total de la population d'une variable de réponse, nous considérons le cas dans lequel la réponse est observée pour tous les éléments de la liste. Les éléments d'une base aréolaire sont échantillonnés avec des probabilités $p_{ij} = 0,05$ et $0,20$. Nous ne considérons pas d'estimations du total de la population fondées sur des sous-échantillons de chaque liste. L'estimation du total de la population, $N_{p_{ij}}$, se présente sous la même forme que (16) avec π_j définie en (15). De même, l'estimation de la taille de la population, $N_{p_{ij}}$, se présente sous la même forme que (9).

L'estimateur

$$\hat{Y}_{CH, p_A} = N_{CH} \bar{Y}^{(p_A)}, p_A = 0,05, 0,20$$

est également considéré, où $\bar{Y}^{(p_A)}$ est la moyenne de l'échantillon des y_j incluses dans l'«échantillon». Le rendement de \hat{Y}_{CH, p_A} dépend de N_{CH} , qui a sous-estimé N apparemment pour le modèle 2. Les résultats pour \hat{Y}_{CH, p_A} ne sont pas inclus ici. Un autre estimateur non biaisé pour le plan de Y est donné par

d'une covariable x_j . Deuxièmement, nous supposons que la covariable peut être corrélée avec la variable de réponse y_j . De plus, nous supposons que x_j et y_j suivent une distribution normale logarithmique avec corrélation ρ_{xy} . La distribution normale logarithmique est utilisée de façon à adapter une distribution asymétrique des covariables. Nous génerons x_j comme e^{u_j} et y_j comme e^{v_j} , où u_j et v_j sont des variables aléatoires normales à deux dimensions avec moyennes nulles, des variances unitaires et une corrélation ρ_{uv} . Il est possible de montrer que $\rho_{uv} = \log[p_{xy} / (e - 1) + 1]$.

Considérons une population de taille N . Supposons qu'il existe deux listes indépendantes, B_1 et B_2 , et une base aréolaire, A . La base aréolaire est supposée complète en ce sens qu'elle couvre la population entière. On tire un échantillon de segments de la base aréolaire et on observe les unités au sein de chaque segment aréolaire. Soit p_{ij} la probabilité d'inclusion d'un élément quelconque dans l'échantillon de la base aréolaire, où p_{ij} est supposée identique pour tous les individus.

La probabilité que le i -ième élément soit inclus dans la j -ième liste est donnée par le modèle de régression logistique (1) pour $i = 1, \dots, N$ et $j = B_1, B_2$. Nous supposons que la probabilité que le i -ième élément soit inclus dans la liste B_1 est indépendante de son état d'inclusion dans la liste B_2 et dans l'échantillon de la base aréolaire.

3.2 Régilage des paramètres

Nous considérons diverses valeurs paramétriques. Pour la taille de la population, N , nous posons $N = 300$ ou $1\,000$. Nous utilisons $\rho_{xy} = -0,3, 0,0, 0,5$ et 1 correspondant à une corrélation négative, nulle, positive et parfaite entre la variable de réponse et la covariable. Ici, $\rho_{xy} = 1$ correspond à $x_i = y_i$, indiquant que la probabilité d'inclusion est liée directement à la variable de réponse.

Pour chacun des $2 \times 4 = 8$ réglages paramétriques ci-dessus de N et ρ_{xy} , nous considérons deux modèles correspondant à différents choix de $\alpha_{B_1}, \beta_{B_1}, \alpha_{B_2}$ et β_{B_2} . Rappelons que $E(x_j) = E[e^{u_j}] = e^{0,5}$. Considérons un élément à valeur de covariable donnée par la valeur moyenne $e^{0,5}$. La probabilité que cet élément soit inclus dans la j -ième liste est

$$p_{j(E)} = \frac{\exp(\alpha_j + \beta_j e^{0,5})}{1 + \exp(\alpha_j + \beta_j e^{0,5})}, j = B_1, B_2.$$

Si $\alpha_j = -\beta_j e^{0,5}$, cet élément a 50% de chances d'être inclus dans la liste j . Nous utilisons cette relation dans le modèle 1.

Nous étendons l'idée ci-dessus, si nous posons

$$\alpha_j = \log\left(-\frac{1}{P} - \frac{P}{1 - P}\right) - \beta_j e^{0,5},$$

alors l'unité à valeur de covariable moyenne comporte une probabilité P d'être incluse dans la liste j . Si nous supposons que les probabilités d'inclusion sont les mêmes pour les listes B_1 et B_2 , alors les chances d'inclusion dans

2.3 Estimation du total de la population à l'aide de listes

Supposons que des valeurs y_i sont disponibles pour tous les éléments de deux listes indépendantes B_1 et B_2 . Pour θ connu, une estimation du total de la population, Y , est l'estimateur de Horvitz-Thompson

$$\hat{Y}^{H-T} = \sum_{i=1}^{M_1} \frac{\pi_i(\theta)}{y_i} \quad (12)$$

D'après Cochran (1977), la variance estimée de \hat{Y}^{H-T} est

$$V(\hat{Y}^{H-T}) = \sum_{i=1}^{M_1} \frac{\pi_i^2(1 - \pi_i(\theta))}{y_i^2(1 - \pi_i(\theta))}.$$

Lorsque θ est inconnu et estimé par $\hat{\theta}$, une estimation du total de la population est

$$\hat{Y}^{H-T} = \sum_{i=1}^{M_1} \frac{\pi_i(\hat{\theta})}{y_i}.$$

Une estimation de la variance de \hat{Y}^{H-T} est dérivée à l'aide de la méthode de Taylor et se présente sous la forme

$$V(\hat{Y}^{H-T}) = \sum_{i=1}^{M_1} \frac{y_i^2(1 - \pi_i(\hat{\theta}))}{\pi_i^2(\hat{\theta})} + B \sum_{i=1}^{M_1} (\hat{\theta}) B_i,$$

$$B = \sum_{i=1}^{M_1} \left[\frac{\pi_i^2(\hat{\theta})}{y_i} \frac{\partial \pi_i(\hat{\theta})}{\partial \hat{\theta}} \right] \quad (13)$$

où

et $\sum(\hat{\theta})$ est l'inverse de la matrice hessienne évaluée sur $\hat{\theta}$. Ces idées s'étendent facilement de façon à incorporer k listes indépendantes.

Dans la pratique, les y_i ne sont pas nécessairement observées pour toutes les unités des listes. Considérons le cas dans lequel les y_i sont disponibles pour un échantillon aléatoire seulement de n_{B_1} et n_{B_2} unités des listes B_1 et B_2 , respectivement. De par leur construction, les probabilités d'inclusion, p_i , varient selon l'individu i et la base j . Toutefois, une fois des individus inclus dans une liste, on en tire un sous-échantillon par échantillonnage aléatoire simple. Par conséquent, tous les éléments de la liste B_j ont des chances égales (n_{B_j}/N_{B_j}) d'être inclus dans le sous-échantillon. À noter que nous tirons des échantillons de chaque liste au lieu de tirer un seul échantillon d'une base de listes combinées. Puisque les listes sont supposées indépendantes, la probabilité estimée que le i -ième individu soit inclus dans au moins une des deux listes est

$$\pi_i = p_{iB_1} \frac{N_{B_1}}{n_{B_1}} + p_{iB_2} \frac{N_{B_2}}{n_{B_2}} - p_{iB_1} p_{iB_2} \frac{N_{B_1}}{n_{B_1}} \frac{N_{B_2}}{n_{B_2}}. \quad (14)$$

On obtient une estimation de Horvitz-Thompson estimée pour Y en substituant (14) en (12).

qui est l'estimateur de Chapman multiplié par la moyenne des réponses pour les éléments inclus dans au moins un sous-échantillon de liste. Encore une fois, cet estimateur est valide uniquement lorsque les probabilités d'inclusion sont homogènes. Il existe $N_{b_1} + N_{b_2} + N_{b_1 b_2}$ éléments uniques dans les bases B_1 et B_2 . Un estimateur semblable se laisse définir lorsque des renseignements sont disponibles uniquement pour des sous-échantillons des listes.

2.4 Estimation du total de la population à l'aide de bases aréolaires et de listes

Considérons le cas dans lequel, en plus de valeurs y_i pour les unités des listes (ou les sous-échantillons des listes), des valeurs y_i sont disponibles pour tous les éléments d'un échantillon aléatoire de segments d'une base aréolaire. L'inclusion de l'information de la base aréolaire donne lieu à la probabilité d'inclusion estimée pour le i -ième individu, c'est-à-dire

$$\pi_i = \pi_i + p_{iA}(1 - \pi_i), \quad (15)$$

où π_i est définie en (4) ou en (14), selon que y_i est observée pour toutes les unités des listes ou seulement pour un sous-échantillon d'unités, respectivement. Dans ce cas, un estimateur de Horvitz-Thompson estimé pour le total de la population est

$$\hat{Y}^{H-T} = \sum_{i \in \text{échantillon}} \frac{\pi_i}{y_i}. \quad (16)$$

Une estimation de la variance de \hat{Y}^{H-T} est donnée par

$$V(\hat{Y}^{H-T}) = \sum_{i=1}^{M_1} \frac{\pi_i^2}{y_i^2(1 - \pi_i)}$$

$$+ 2 \sum_{i=1}^I \sum_{j=1}^I \frac{\pi_i \pi_j \pi_{ij}}{(\pi_i - \pi_i \pi_j)(\pi_j - \pi_j \pi_i)} y_i y_j + B \sum_{i=1}^I B_i, \quad (17)$$

où π_{ij} est définie en (11) et B et \sum sont définis en (13).

3. ÉTUDE DE SIMULATION

3.1 Hypothèses de l'étude

À fin d'étudier les propriétés de la taille de la population et des estimateurs du total, Haines (1997) a considéré quatre-vingt modèles différents. Nous présentons ici des détails pour deux de ces modèles seulement. Il est supposé que les probabilités d'inclusion pour deux listes dépendent

2.2 Estimation de la taille de la population à l'aide de listes et de bases aréolaires

Supposons que nous ayons accès à une base aréolaire en plus des deux listes B_1 et B_2 . La base aréolaire est constituée de U_4 segments qui couvrent la population entière. Un échantillon aléatoire simple de u_4 segments est sélectionné. Nous supposons que toutes les unités des segments échantillonnés sont observées. La probabilité d'inclusion dans l'échantillon de la base aréolaire est la même pour toutes les unités et représentée la quantité connue $P_4 = u_4/U_4$. Ensuite, nous maximisons la vraisemblance conditionnelle (3) relativement à θ et nous calculons la probabilité estimée qu'un individu i soit inclus dans au moins une liste ou dans la base aréolaire. Cette probabilité est notée $\pi_i = \pi_i + P_4(1 - \pi_i)$. Les probabilités π_i et \hat{p}_i sont définies en (4) et en (5), respectivement. Un estimateur de Horvitz-Thompson estime pour la taille de la population est

$$\hat{N} = \sum_{i \in \text{échantillon}} \frac{\pi_i}{1} \quad (9)$$

Cet estimateur se laisse facilement étendre au cas comportant k listes, B_1, \dots, B_k , et une base aréolaire indépendante. D'après Cochran (1977), une estimation de la variance de \hat{N} est donnée par

$$V(\hat{N}) = \sum_{i=1}^{M_1} \frac{1}{1 - \pi_i} + 2 \sum_{i < j}^i \frac{\pi_i \pi_j}{(\pi_i - \pi_j)(\pi_j - \pi_i)} + \sum \hat{A}_i, \quad (10)$$

où \hat{A} est défini en (7) et \sum est l'inverse de la matrice hessienne. La formule de variance pour \hat{N} en (6) et son estimation sont valides uniquement lorsque π_{ij} , la probabilité que les unités i et j soient incluses dans l'échantillon, est égale à $\pi_i \pi_j$. Lorsqu'on inclut un échantillon de base aréolaire, toutefois, π_{ij} n'est pas nécessairement égale à $\pi_i \pi_j$. Supposons que les unités i et j appartiennent au même segment de base aréolaire. Dans un tel cas, les unités i et j sont toutes deux incluses ou non incluses dans l'échantillon, selon que leur segment correspondant est sélectionné ou non. Il est possible de montrer que la probabilité d'inclusion combinée, π_{ij} , se laisse estimer comme suit:

$$\pi_{ij} = \begin{cases} \pi_i \pi_j & \text{si } i \text{ et } j \text{ appartiennent à des segments aréolaires différents} \\ \pi_i + \pi_j - P_4 & \text{si } i \text{ et } j \text{ appartiennent au même segment aréolaire} \end{cases} \quad (11)$$

où π_i est définie en (4) et $\pi_i = \pi_i + P_4(1 - \pi_i)$. Ainsi, lorsque i et j appartiennent au même segment, $\pi_{ij} \neq \pi_i \pi_j$. Toutefois, si P_4 est petite et que π_i et π_j sont grandes, alors $(\pi_{ij} - \pi_i \pi_j)$ se rapproche de zéro.

$$V(\hat{N}) = \sum_{i=1}^{M_1} \frac{1}{1 - \pi_i(\theta)} \pi_i^2(\theta) \hat{A}_i + A \hat{\Sigma}(\theta) \hat{A},$$

où

$$\hat{A} = \sum_{i=1}^{M_1} \left[\frac{1}{\partial \pi_i(\theta)} \pi_i^2(\theta) \frac{\partial \theta}{\partial \theta} \right]$$

et $\hat{\Sigma}(\theta)$ est l'inverse de la matrice hessienne. Le second membre en (7) est attribuable à l'estimation de $\pi_i(\theta)$ par un autre estimateur de la taille de la population qui est utilisé couramment dans des expériences de type saisir-ressaier est l'estimateur de Lincoln-Petersen. Cet estimateur classique est attribuable à Lincoln (1930) et à Petersen (1896) et se présente comme suit:

$$\hat{N}^{L-P} = \frac{N_{B_1} N_{B_2}}{N_{B_1 B_2}},$$

où N_{B_1} et N_{B_2} désignent la taille des listes B_1 et B_2 , respectivement, et $N_{B_1 B_2}$ désigne le nombre d'unités communes aux deux bases de sondage. Il s'agit d'une méthode simple d'estimation de moments fondée sur l'hypothèse voulant que toutes les unités comportent des probabilités d'inclusion homogènes pour chacune des deux listes indépendantes. Il se peut que le dénominateur $N_{B_1 B_2}$ soit nul. Chapman (1951) a proposé une version modifiée de l'estimateur de Lincoln-Petersen, donnée par

$$\hat{N}^{CH} = \frac{(N_{B_1} + 1)(N_{B_2} + 1)}{(N_{B_1 B_2} + 1)} - 1. \quad (8)$$

Cet estimateur est moins biaisé que l'estimateur de Lincoln-Petersen (Chapman 1951). D'après Sekar et Deming (1949), l'erreur-type asymptotique de \hat{N}^{CH} est

$$\sqrt{V(\hat{N}^{CH})} = \sqrt{\frac{N_{B_1} N_{B_2} N_{B_1 B_2}}{(N_{B_1 B_2})^3}}.$$

où N_{B_1} et N_{B_2} désignent le nombre d'unités appartenant uniquement aux listes B_1 et B_2 , respectivement.

L'estimateur de Lincoln-Petersen est l'estimateur du maximum de vraisemblance inconditionnelle de la taille de la population N lorsqu'il y a deux listes indépendantes et que les probabilités d'inclusion sont homogènes. Haines (1997) étend les procédures d'estimation à k listes, chacune comportant des probabilités d'inclusion homogènes. Cet estimateur, toutefois, n'est pas approprié lorsque les probabilités d'inclusion sont hétérogènes. On trouvera à la section 3 les résultats de la simulation.

À part la stratégie non paramétrique, une autre façon de procéder est de modéliser les probabilités d'inclusion en fonction d'une variable auxiliaire. Pollock, Hines et Nichols (1984), Huggins (1989) et Alho (1990) ont examiné le rôle des variables auxiliaires dans des expériences de type (inclusion). Les expériences de type saisir-ressaisir pour des populations fermées comportent $i = 1, \dots, N$ individus et $j = 1, \dots, t$ cycles de piégeage. Encore une fois, les $j = 1, \dots, t$ cycles de piégeage sont semblables à $t = k$, le nombre de listes indépendantes. Huggins (1989) et Alho (1990) ont proposé une procédure d'estimation conditionnelle en vue de l'estimation de la taille d'une population fermée se fondant sur une saisie et une seule ressaie. Dans ces deux exposés, les auteurs ont adopté le modèle logistique pour les probabilités d'inclusion, données par

$$(1) \quad p_{ij} = \frac{\exp(\alpha_j + \beta_j x_i)}{1 + \exp(\alpha_j + \beta_j x_i)},$$

où x_i est une covariable et α_j et β_j sont des paramètres inconnus. À noter que cette paramétrisation donne $0 \leq p_{ij} \leq 1$ pour toutes les valeurs de α_j et β_j . Pour $\beta_j > 0$, la probabilité d'inclusion augmente en fonction de la covariable. Cette paramétrisation est différente de l'échantillonnage supposée proportionnelle à x_i . Les estimateurs du maximum de vraisemblance de α_j et β_j sont obtenus en fonction de la vraisemblance conditionnelle que l'unité se retrouve dans une liste au moins. Haines (1997) a dérivé la fonction de vraisemblance conditionnelle pour trois listes indépendantes.

En traitant chaque individu comme une strate distincte, définissons les variables indicatrices ci-dessous pour $i = 1, \dots, N$:

$$n_{ij} = \begin{cases} 1 & \text{l'individu } i \text{ est compris dans la base de sondage } j \text{ seulement} \\ 0 & \text{autrement} \end{cases}$$

et

$$j = B_1, B_2, \dots$$

$$a_i = \begin{cases} 1 & \text{l'individu } i \text{ est compris dans les deux bases de sondage} \\ 0 & \text{autrement} \end{cases}$$

La valeur de l'expression

$$M'_i = n_{iB_1} + n_{iB_2} + a_i$$

(2) est 1 si l'individu i est compris dans au moins une des deux bases de sondage et 0 autrement.

Alho (1990) a présenté la fonction de vraisemblance conditionnelle pour deux listes comme suit:

$$(3) \quad \frac{\exp\{\alpha_{B_1} N_{B_1} + \alpha_{B_2} N_{B_2} + \beta_{B_1} \sum_{i \in B_1} x_i + \beta_{B_2} \sum_{i \in B_2} x_i\}}{\prod_{M'_i=1}^N K'_i(\theta)},$$

où

$$K'_i(\theta) = \exp\{\alpha_{B_1} + \beta_{B_1} x_i\} + \exp\{\alpha_{B_2} + \beta_{B_2} x_i\} + \exp\{\alpha_{B_1} + \alpha_{B_2} + (\beta_{B_1} + \beta_{B_2}) x_i\}$$

et $\theta = (\alpha_{B_1}, \beta_{B_1}, \alpha_{B_2}, \beta_{B_2})$. Alho (1990) a utilisé une procédure itérative fondée sur les statistiques suffisantes pour maximiser (3) et nous appliquons la méthode de Newton pour calculer des estimateurs du maximum de vraisemblance conditionnelle pour θ , notés $\hat{\theta} = (\hat{\alpha}_{B_1}, \hat{\beta}_{B_1}, \hat{\alpha}_{B_2}, \hat{\beta}_{B_2})$. On trouvera à l'annexe A dans Haines (1997) des détails sur la méthode de Newton. La probabilité estimée que l'individu i soit compris dans au moins une des listes est notée

$$(4) \quad \pi_i = 1 - \left(\frac{1}{1 + \exp(\hat{\alpha}_{B_1} + \hat{\beta}_{B_1} x_i)} \right) \times \left(\frac{1}{1 + \exp(\hat{\alpha}_{B_2} + \hat{\beta}_{B_2} x_i)} \right) = \pi_i(\hat{\theta}),$$

$$p_{ij} = \frac{\exp(\hat{\alpha}_j + \hat{\beta}_j x_i)}{\exp(\hat{\alpha}_j + \hat{\beta}_j x_i) + 1 + \exp(\hat{\alpha}_j + \hat{\beta}_j x_i)},$$

(5) Si θ était connu, l'estimateur de Horvitz-Thompson de N serait $\hat{N} = \sum_{M'_i=1}^N 1/\pi_i$ (Horvitz et Thompson 1952). D'après Cochran (1977), la variance de \hat{N} est

$$(6) \quad V(\hat{N}) = \sum_{i=1}^N \frac{\pi_i}{1 - \pi_i}.$$

Une estimation de la variance de \hat{N} est

$$\hat{V}(\hat{N}) = \sum_{M'_i=1}^N \frac{\pi_i^2}{1 - \pi_i}.$$

Puisque θ est inconnu, nous considérons l'estimation de la taille de la population donnée par $\hat{N} = \sum_{M'_i=1}^N 1/\hat{\pi}_i$, où $\hat{\pi}_i$ est définie en (4). Une estimation de la variance de \hat{N} est obtenue à l'aide de la méthode de Taylor et se présente sous la forme

2. PROBABILITÉS D'INCLUSION
HÉTÉROSCÉDASTIQUES

2.1 Estimation de la taille de la population à l'aide
de listes

Dans des expériences de type saisir-ressaisir, il se peut que différents animaux comportent différentes probabilités de saisie. De même, des éléments individuels peuvent porter différentes probabilités d'inclusion dans une base de sondage sous forme de liste. Des listes différentes peuvent être considérées comme des cycles de saisie différents. Le modèle M_h désigne le modèle d'hétérogénéité dans la documentation sur les démarches de type saisir-ressaisir pour une population fermée (Otis, Burnham, White et Anderson 1978). Dans un contexte de type saisir-ressaisir, les probabilités de saisie, bien que pouvant varier d'un animal à l'autre, sont supposées les mêmes pour tous les cycles de piégeage. Le modèle d'hétérogénéité peut comporter jusqu'à $N + 1$ paramètres au total, c'est-à-dire N et $p_i, i = 1, \dots, N$, où N est la taille de la population et p_i la probabilité d'inclusion pour la i -ième unité. Pour désigner la probabilité d'inclusion pour la i -ième unité, laquelle la probabilité d'inclusion p_i pour l'élément i est constante pour l'ensemble des k listes, B_1, B_2, \dots, B_k .

Burnham (1972) et Burnham et Overton (1978, 1979) ont examiné le problème de l'estimation de N dans un contexte de type saisir-ressaisir. L'estimateur proposé pour N que décrit Burnham (1972) se fonde sur la méthode jackknife de réduction du biais (Quenouille 1956). Chao (1988) a élaboré un autre estimateur de moment pour ce modèle en fonction de données sur la fréquence de saisie (Pollock 1991). Dans certaines circonstances, l'estimateur proposé par Chao est moins biaisé que l'estimateur jackknife de Burnham. En général, il est difficile de trouver un estimateur tout à fait satisfaisant de N dans le cadre du modèle M_h . Otis et coll. (1978) ont donc suggéré que l'on conçoive l'étude dans son ensemble de façon à minimiser l'hétérogénéité. Norris et Pollock (1996) ont proposé un estimateur du maximum de vraisemblance non paramétrique qui n'est pas tout à fait satisfaisant.

Dans des expériences de type saisir-ressaisir, le modèle exprimé sous forme du M -ième modèle admet des probabilités d'inclusion pouvant varier tant selon le cycle de piégeage (base sous forme de liste) que selon l'individu. Nous définissons p_{ij} comme la probabilité d'inclusion du i -ième élément de la j -ième liste. Le M -ième modèle n'est évidemment pas facile à estimer puisqu'il peut comporter jusqu'à $(N + 1)$ paramètres, où $i = k$, le nombre de listes. Chao, Lee et Jeng (1992), employant l'idée de couverture de l'échantillon, ont proposé une méthode non paramétrique d'estimation de la taille de la population pour le M -ième modèle.

Une solution modélisée de type saisir-ressaisir en vue de l'estimation de la taille des bases de sondage en fonction d'informations tirées de deux listes incomplètes. D'après Cochran (1977), il est souvent difficile d'obtenir une liste qui corresponde exactement à la population étudiée. Les listes que l'on établit habituellement pour un but quelconque sont souvent incomplètes et en partie illisibles, ou comportent une part inconnue de dédoublement. Puisque les listes sont en règle générale incomplètes, les estimations qui se fondent uniquement sur des listes risquent de sous-estimer la taille de la population. En complétant les renseignements disponibles à l'aide d'un échantillon tiré d'une base aréolaire, on pourra peut-être établir des estimations efficaces de la taille de la population et des chiffres de population.

Une base aréolaire est une collection de régions géographiques définies par des limites identifiables. Les chargés d'enquête utilisent souvent des bases aréolaires afin d'obtenir une couverture complète de la population cible. Des populations comme les exploitations agricoles sont naturellement associées aux subdivisions de terrain constituant une base aréolaire. Ainsi, dans une enquête agricole, la région d'intérêt est divisée en une série de régions géographiques appelées des segments. Les segments, qui sont les unités d'échantillonnage, sont alors sélectionnés à l'aide de plans de sondage à plusieurs degrés stratifiés (Kott et Vogel 1995). On définit des règles qui établissent les liens entre les exploitations agricoles de la population et les segments de la base aréolaire. Lorsque les exploitations agricoles, ou les unités déclarantes, ont été identifiées au sein de chaque segment d'échantillonnage, on les dénombre en personne et on recueille les données pertinentes. Nealon (1984) a donné une description détaillée des estimateurs de segments ouverts, fermés et pondérés. Faulkenberry et Garoui (1991) ont formulé des estimateurs supplémentaires conçus précisément pour des bases aréolaires. On trouvera dans Fecso et coll. (1986) des méthodes plus complexes de construction et d'échantillonnage pour les bases aréolaires. L'échantillonnage et le sous-échantillonnage à partir de bases aréolaires sont examinés en détail dans Kott et Vogel (1995). À la section 2, les auteurs examinent des listes indépendantes dont les éléments comportent des probabilités d'inclusion hétéroscédastiques. Il y est question de méthodes qui fournissent des estimateurs de la taille de la population et des chiffres de population en présence de renseignements tirés d'une ou plusieurs listes et d'une base aréolaire. La section 3 résume les résultats d'une étude de simulation qui a permis de comparer divers estimateurs de la taille de la base (population) et des chiffres de population. Enfin, les auteurs présentent des conclusions et un sommaire.

Estimation de la taille et des chiffres de population pour des échantillons tirés de listes incomplètes avec probabilités d'inclusion hétérogènes

DAWN E. HAINES, KENNETH H. POLLOCK et SASTRY G. PANTULA¹

RÉSUMÉ

Les informations tirées de bases de sondage aréolaires et de listes sont combinées de façon à fournir des estimations efficaces de la taille et des chiffres de population. Les auteurs examinent le cas où les probabilités d'inclusion dans les listes sont hétérogènes et modélisées en fonction de covariables. Ils adaptent et modifient la méthode employée par Huggins (1989) et par Alho (1990) pour la modélisation de variables auxiliaires dans des études de type saisi-r-ressaisi faisant appel à un modèle de régression logistique. Les auteurs présentent les résultats d'une étude de simulation qui permet de comparer divers estimateurs de la taille des bases de sondage et des chiffres de population en ayant recours à la stratégie de régression logistique pour modéliser des probabilités d'inclusion hétérogènes.

MOTS CLÉS : Régression logistique; base de sondage sous forme de liste; base de sondage aréolaire; échantillonnage de type saisi-r-ressaisi.

1. INTRODUCTION

Dans le présent exposé, les auteurs estiment la taille de la population et les chiffres de population en présence d'informations tirées de plusieurs bases de sondage indépendantes. Il est supposé que les éléments de la population comportent différentes probabilités d'inclusion pour diverses bases de sondage. Ces probabilités d'inclusion hétérogènes peuvent dépendre d'une covariable. Ainsi, supposons qu'il s'agisse d'estimer le nombre d'exploitations porcines et le nombre de cochons en Caroline du Nord. Des mesures de covariable comme la superficie des exploitations porcines ou le nombre d'employés sont une indication de la taille des exploitations porcines. Il se peut que de grandes exploitations aient de meilleures chances d'être incluses dans une liste servant de base de sondage que des exploitations plus petites. Dans des expériences de type saisi-r-ressaisi, il se peut que les probabilités de saisie soient inégales pour les animaux. Les probabilités de saisie (inclusion) pour les animaux peuvent varier selon l'âge, le sexe, la taille ou l'espèce.

On entend par listes de bases de sondage qui énumèrent des unités d'échantillonnage de la population cible. Les éléments qu'on y trouve peuvent comprendre, entre autres, des noms, des adresses, des numéros de téléphone, des numéros de sécurité sociale ou des descriptions matérielles d'emplacement. Ces variables et d'autres variables de stratification diverses servent à identifier des personnes, des animaux, des entreprises ou d'autres établissements. Les listes et les bases aréolaires servent à fournir des estimations de taille et de chiffres de population inconnus. Puisque, dans toute importante opération de collecte de données, les bases de sondage comportent inévitablement des imperfections

comme des omissions, des doublons et des enregistrements inexacts (Hansen, Hurwitz et Madow 1953), on trouve dans la littérature diverses propositions de solution pour surmonter cette difficulté. Une stratégie, élaborée d'abord par Hartley (1962, 1974), combine une liste incomplète et une base aréolaire. D'autres développements théoriques sont dus à Cochran (1965), à Lund (1968), à Fuller et Burneister (1972) et à Bosecker et Ford (1976). Haines et Pollock (1998a) ont appliqué la méthode des deux bases de sondage à une population de pigargues à tête blanche, tandis que Haines et Pollock (1998b) ont présenté une stratégie théorique plus générale pour ce qui est de la combinaison de plusieurs bases de sondage. Ces deux exposés n'ont pas examiné le cas des probabilités d'inclusion hétérogènes. Fienberg (1992) a présenté une bibliographie annotée de la documentation sur les démarches de type saisi-r-ressaisi traitant plus particulièrement du problème de sous-dénombrement du recensement, qui comprend les exposés de Wolter (1986, 1990) et de Cowan et Malec (1986).

Le NASS (*National Agricultural Statistics Service*) utilise actuellement plusieurs bases de sondage pour ses démarches d'échantillonnage et d'estimation de nombreux produits agricoles. Le NASS recueille et résume des données sur les superficies ensemencées, le bétail, la production et les stocks de céréales, les coûts de production, les dépenses agricoles et d'autres caractéristiques agricoles. Fecso, Tortora et Vogel (1986) ont passé en revue les bases de sondage utilisées pour le secteur agricole des États-Unis, tandis que Nealson (1984) a fourni des détails sur les estimateurs liés à des bases de sondage multiples et aréolaires qui sont utilisés par le Département of Agriculture des États-Unis. Pollock, Turner et Brown (1994) ont présenté

¹ Dawn E. Haines, U. S. Bureau of the Census, Washington, DC 20233; Kenneth H. Pollock et Sasthy G. Pantula, North Carolina State University, Department of Statistics, Box 8203, Raleigh, NC 27695-8203, U.S.A.

méritées. Il s'est vu décerner un doctorat honoraire de l'Université de Bologne à l'occasion du 900^e anniversaire de l'établissement, la médaille Samuel Wilks, soit la plus haute distinction de l'ASA, la chaire Henry Russell, le titre plus haute distinction de l'Université du Michigan, le titre de membre titulaire honoraire de l'IISS que je considère comme une forme de prix Nobel de la statistique et peut-être la reconnaissance la plus significative à ses yeux, sans compter la multitude de hautes distinctions conférées par la Hongrie (doctorat honoraire de la plus grande université de Budapest, titre de membre honoraire de l'Académie des sciences de la Hongrie et la Croix d'officier de l'Ordre du Mérite).

Outre ce qu'il nous légua dans le domaine de la statistique, il nous laisse le phénomène «Leslie Kish, une force de la nature»: le combattant pendant la Guerre civile d'Espagne, le philosophe de tout ce qui est statistique, le défenseur des droits humains qui n'aura jamais vieilli, le conteur, le lecteur insatiable, l'auteur des plus belles lettres de Noël, l'époux et le père affectueux et l'ami de toujours de certaines et peut-être de milliers de personnes.

Lorsque j'ai pris la parole à son 90^e anniversaire de naissance, j'ai terminé en disant espérer être présent au véritable anniversaire de Leslie, celui que la Reine-Mère venait de célébrer. Et je ne plaisantais pas. Il était si plein de vie qu'il semblait non seulement possible de le voir atteindre le centenaire, mais en fait il paraissait même impossible d'imaginer autre chose. Malheureusement, il s'est éteint. Son dernier acte de générosité aura consisté à offrir son corps à la recherche médicale. Ne serait-il pas pertinent que les travaux qui en résultent puissent nous permettre de mieux comprendre cet être exceptionnel que fut Leslie Kish?

s'intéresse, dès le début, aux estimations visant les petites régions; et j'en passe. Mais, sans vouloir déprécier l'importance de ces travaux, j'estime que certaines de ses autres contributions sont tout aussi fondamentales.

Il fut l'une des rares personnes dont les premiers travaux appliqués ont conféré à l'échantillonnage respect et admiration. En plus de compter au nombre des fondateurs de ce qui est devenu l'*Institute for Survey Research* à Ann Arbor, il enseigne à des générations de statisticiens, tant américains qu'étrangers, dans le cadre du légendaire programme d'été pour les statisticiens étrangers. Après avoir officiellement pris sa retraite, il poursuit ses activités par les cours qu'il donne dans le programme d'été, par le travail d'édition et de collaboration à l'une ou l'autre des chroniques de questions et de réponses du bulletin *The Survey Statistician* entrepris depuis des décennies ainsi que par une multitude de conférences et travaux de services de conseils. Dans les rencontres internationales, il m'arrivait de «tomber» sur ses anciens étudiants et amis du temps. Aujourd'hui, on ne «tombe» plus sur eux par hasard, ils sont littéralement omniprésents: je me demande combien de spécialistes étrangers de l'échantillonnage bien connus n'ont pas été à un moment donné des étudiants de Leslie Kish. Et je ne veux pas passer sous silence deux de ces nombreuses contributions qui me tiennent particulièrement à cœur: ses années de loyaux services à Statistique Canada en qualité de membre fondateur de notre Comité consultatif des méthodes statistiques et son allocation à titre de président de l'ASA en 1977 (laquelle a paru dans le *JASA* en mars 1978), l'allocation la plus marquante du président de l'ASA qu'il m'a été donné

d'entendre. Ses réalisations lui ont valu une reconnaissance à l'échelle internationale. Je me contenterai de souligner quelques-unes des douzaines de distinctions qu'il s'est

Leslie Kish – Une vie de dévouement

IVAN P. FELLEGI¹

1. INTRODUCTION

Je ne peux croire qu'il me faut écrire un article à la mémoire de Leslie Kish. Il y a tout juste quelques mois, j'ai rédigé une courte allocution quelque peu humoristique à l'occasion de son 90^e anniversaire de naissance. J'avais alors demandé en plaisantant pourquoi on faisait tant de cas d'un 90^e anniversaire. Après tout, la Reine-Mère venait à peine de célébrer son centenaire. J'avais alors remarqué que *cela* méritait d'être souligné. Il avait ri de bon cœur avec, dans les yeux, ce petit larmier bien connu qui lui était propre. J'étais de nouveau surpris de constater combien il aimait encore s'amuser, la force de son dynamisme, de sa perspicacité et, en fait, de sa jeunesse dans toutes les facettes de son comportement et ce, malgré une mobilité quelque peu réduite. Il m'avait parlé de l'arthroplastie partielle du genou qu'il devait subir et m'avait confié que, selon son médecin, il devait choisir entre l'opération ou la marchette pour se déplacer. Bien entendu, il n'était pas question pour lui d'opter pour la marchette: il lui fallait jouir d'une mobilité complète. Et cette mobilité, à 90 ans, ne signifiait pas simplement la capacité de se déplacer à la maison; il s'agissait plutôt de voyager partout dans le monde plusieurs fois par année. Leslie Kish est mort des suites de complications postopératoires, après avoir lutté plusieurs semaines avec le courage inébranlable qui le caractérisait.

À mes yeux, le trait le plus distinctif de cet homme fut son dévouement inaltérable. L'un de ses derniers actes d'altruisme a été d'inspirer ses amis et collègues à établir le *Leslie Kish International Fellows Fund* dans le but d'aider les étudiants de pays en développement à obtenir une formation en échantillonnage de populations.

Leslie Kish est né en 1910 à Poprad, qui faisait alors partie de l'Empire austro-hongrois et est maintenant située en Slovaquie. Il racontait que, à différentes époques de l'histoire, Poprad appartenait à cinq pays différents; c'est là un symbole éloquent d'une vie motivée par l'amour des citoyens de toutes les régions du monde. En 1925, ses parents décident d'émigrer aux États-Unis, à l'instar de centaines de milliers de Hongrois quittant leur pays. Comme l'a dit le célèbre poète hongrois, Attila József, «un million de demi de nos citoyens sont sortis du pays en titubant pour gagner l'Amérique». Peu après leur arrivée, le père de Leslie décède. La mère de Leslie et les quatre enfants doivent décider s'ils resteront aux États-Unis. C'est ce qu'ils choisissent de faire, mais cela signifie que les deux aînés, dont Leslie alors âgé de 16 ans, devront travailler pour subvenir aux besoins de la famille.

Leslie Kish ne peut croire qu'il me faut écrire un article à la mémoire de Leslie Kish. Il y a tout juste quelques mois, j'ai rédigé une courte allocution quelque peu humoristique à l'occasion de son 90^e anniversaire de naissance. J'avais alors demandé en plaisantant pourquoi on faisait tant de cas d'un 90^e anniversaire. Après tout, la Reine-Mère venait à peine de célébrer son centenaire. J'avais alors remarqué que *cela* méritait d'être souligné. Il avait ri de bon cœur avec, dans les yeux, ce petit larmier bien connu qui lui était propre. J'étais de nouveau surpris de constater combien il aimait encore s'amuser, la force de son dynamisme, de sa perspicacité et, en fait, de sa jeunesse dans toutes les facettes de son comportement et ce, malgré une mobilité quelque peu réduite. Il m'avait parlé de l'arthroplastie partielle du genou qu'il devait subir et m'avait confié que, selon son médecin, il devait choisir entre l'opération ou la marchette pour se déplacer. Bien entendu, il n'était pas question pour lui d'opter pour la marchette: il lui fallait jouir d'une mobilité complète. Et cette mobilité, à 90 ans, ne signifiait pas simplement la capacité de se déplacer à la maison; il s'agissait plutôt de voyager partout dans le monde plusieurs fois par année. Leslie Kish est mort des suites de complications postopératoires, après avoir lutté plusieurs semaines avec le courage inébranlable qui le caractérisait.

À mes yeux, le trait le plus distinctif de cet homme fut son dévouement inaltérable. L'un de ses derniers actes d'altruisme a été d'inspirer ses amis et collègues à établir le *Leslie Kish International Fellows Fund* dans le but d'aider les étudiants de pays en développement à obtenir une formation en échantillonnage de populations.

Leslie Kish est né en 1910 à Poprad, qui faisait alors partie de l'Empire austro-hongrois et est maintenant située en Slovaquie. Il racontait que, à différentes époques de l'histoire, Poprad appartenait à cinq pays différents; c'est là un symbole éloquent d'une vie motivée par l'amour des citoyens de toutes les régions du monde. En 1925, ses parents décident d'émigrer aux États-Unis, à l'instar de centaines de milliers de Hongrois quittant leur pays. Comme l'a dit le célèbre poète hongrois, Attila József, «un million de demi de nos citoyens sont sortis du pays en titubant pour gagner l'Amérique». Peu après leur arrivée, le père de Leslie décède. La mère de Leslie et les quatre enfants doivent décider s'ils resteront aux États-Unis. C'est ce qu'ils choisissent de faire, mais cela signifie que les deux aînés, dont Leslie alors âgé de 16 ans, devront travailler pour subvenir aux besoins de la famille.

Il formule le concept des plans à objectifs multiples; il explore la question de l'inférence dans les échantillons complexes et met au point le concept novateur désigné aujourd'hui par le terme répétition compensée – *balanced repeated replication* (en collaboration avec Marty Frankel); il fait oeuvre de pionnier dans l'étude de l'erreur de réponse; il se fait l'apôtre des recensements et échantillons avec renouvellement; il lance le concept du choix contrôlé; il formule le concept des plans à objectifs multiples; il

You et Rao présentent des modèles hiérarchiques de Bayes à plusieurs niveaux pour les estimations régionales. Les modèles admettent des paramètres de régression aléatoires qui dépendent également de covariables régionales. La moyenne régionale est estimée à l'aide de la moyenne a posteriori, et la variance a posteriori est prise comme mesure de la précision. Trois modèles de variance sont examinés: fixe égal, fixe inégal, aléatoire. Les auteurs présentent des détails sur l'échantillonnage de Gibbs pour ces modèles et s'en servent comme outil d'inférence. Ils illustrent les procédures à l'aide de données sur le revenu des ménages au Brésil au niveau des comtés.

Okafor et Lee considèrent un schéma d'échantillonnage double, comportant un sous-échantillon de non-répondants qui sont interrogés de nouveau au cours de la deuxième phase en fonction d'un taux d'échantillonnage fixe. Suivant ce schéma, les auteurs proposent des versions modifiées des estimateurs par quotient et par régression. Ils déterminent des valeurs optimales pour la taille des échantillons et le taux d'échantillonnage fixe, d'après des fonctions de coûts, de façon à minimiser la variance. De plus, ils présentent les variances et leurs estimateurs. Une petite étude empirique permet d'étudier l'efficacité relative des estimateurs modifiés par quotient et par régression en fonction de l'estimateur standard de Hansen-Hurwitz.

Pickery et Loosveldt intègrent une importante technique analytique à l'étude de la non-réponse partielle. Leurs modèles offrent un aperçu plus complet des facteurs qui influencent la non-réponse partielle que ne le font les travaux antérieurs. Un aspect important de la stratégie est que les auteurs établissent une distinction entre la variation propre à l'intervieweur ou au répondant, la variation attribuable à des caractéristiques de l'intervieweur ou du répondant et la variance de l'erreur.

Fuchs examine l'influence que la conception de l'écran et l'ordre des questions exercent sur le comportement des intervieweurs dans le cadre d'une interview assistée par ordinateur (IAO). À l'aide d'expériences de laboratoire, on a pu montrer que la conception de l'écran et l'ordre des questions influencent bel et bien le comportement des intervieweurs. Dans son exposé, Fuchs présente les résultats d'une expérience sur le terrain dans laquelle on a analysé deux conceptions de l'écran ainsi que deux ordres des questions dans un plan factoriel 2×2 . Les résultats se fondent sur des mesures du temps intégrées à l'application ITAO et sur 234 interviews choisies au hasard et enregistrées sur bande magnétique puis analysées en fonction d'un schéma de codage.

M.P. Singh

Dans ce numéro

Le présent numéro est dédié à Leslie Kish, qui est décédé cet automne à l'âge de 90 ans. Chose remarquable, jusqu'à la fin de sa vie le professeur Kish a continué de présenter de nouvelles idées statistiques et méthodologiques, comme en témoigne son article sur le cumul et la combinaison des enquêtes démographiques publié il y a un an dans le numéro marquant le 25^e anniversaire de *Techniques d'enquête*. On trouvera au début du présent numéro une réflexion d'Ivan Fellegi sur la vie et les réalisations statistiques du professeur Kish.

L'article de Haines, Pollock et Pantula examine les estimations d'un total lorsque l'on combine deux listes incomplètes et une base aréolaire. Les auteurs proposent des chiffres de population appropriés pour rendre compte du caractère incomplet des bases de sondage. De plus, leurs modèles tiennent compte du fait que les listes incomplètes ont de meilleures chances d'inclure des unités d'échantillonnage plus grandes.

Beaumont propose une méthode d'estimation permettant de réduire le biais dû à un mécanisme de réponse qui dépend de la variable d'intérêt, qu'on appelle un mécanisme de réponse non-ignorable. La méthode proposée requiert un modèle pour la variable d'intérêt et un modèle pour la probabilité de répondre. La méthode est considérée robuste par rapport à l'hypothèse de normalité puisqu'elle est construite de telle sorte qu'elle n'exige pas de spécifier la distribution des erreurs du modèle impliquant la variable d'intérêt, ce qui n'est pas le cas de la méthode du maximum de vraisemblance. Il propose également une méthode simple permettant de vérifier la validité de l'hypothèse de normalité des erreurs quand la non-réponse n'est pas ignorable. Spencer aborde le problème de l'estimation des effets de plan de sondage attribuables à la pondération lorsqu'il existe une corrélation entre les probabilités de sélection et la variable d'intérêt. À l'aide d'une représentation de la population par régression, Spencer présente une approximation de l'effet de plan de sondage lorsqu'il existe une corrélation entre les probabilités de sélection et la variable d'intérêt.

Bienner et Bushney ont recours à l'hypothèse de Markov sur les changements de situation vis-à-vis de l'activité afin de corriger les erreurs de classification dans des données sur la population active. À l'aide de cette méthode, ils estiment les taux d'erreur de réponses pour des panels fournissant des données mensuelles sur la population active dans le cadre de la Current Population Survey (CPS). La cohérence globale des résultats est considérée comme un indicateur de l'utilité de l'analyse markovienne de la structure latente comme moyen d'évaluer l'exactitude des réponses à la CPS. Un aspect critique de cette analyse est la confirmation de l'hypothèse de Markov; les auteurs présentent des données empiriques intéressantes sur sa validité à court terme relatif à la CPS.

De nombreux bureaux de la statistique font appel à la méthode de répétition MHS (demi-échantillon modifié) pour l'estimation de la variance d'échantillonnage des médianes. C'est là un problème concret important puisque le calcul direct des médianes d'échantillon exige souvent une énorme capacité de calcul. Une autre méthode d'estimation consiste à grouper les données continues en intervalles discrets et à utiliser une interpolation linéaire pour l'intervalle comportant la médiane. Dans leur article, Thompson et Sigman comparent les effets d'une absence de groupement (c'est-à-dire la médiane de l'échantillon), d'un groupement avec intervalles de taille fixe et d'un groupement avec des intervalles dont la taille dépend des données, sur les médianes et les estimations connexes de la variance MHS. Leur étude empirique indique que les intervalles dont la taille dépend des données fournissent les estimations de la variance qui comportent le moins de biais, la meilleure stabilité et les meilleurs intervalles de confiance.

McLaren et Steel examinent les répercussions de différents modèles de chevauchement sur la variance d'échantillonnage des estimations désaisonnalisées et des estimations de tendance obtenues de séries chronologiques fondées sur des sondages lorsqu'on utilise les méthodes de désaisonnalisation X-11 et X-11-ARIMA du Recensement. Les profils de renouvellement «dans pour 8», «dans pour 6», «dans pour 4, hors pour 4, dans pour 4» sont raisonnables si la statistique clé à analyser est le changement mensuel des estimations désaisonnalisées. Toutefois, si les statistiques clés sont le niveau tendanciel et la différence entre deux estimations de tendances consécutives, le profil de renouvellement «dans pour 1, hors pour 2, dans pour 1, pour un total de 8 mois» est préférable lorsque l'on veut réduire la variance de l'échantillonnage. Les auteurs indiquent également que le profil «dans pour 2, hors pour 2, dans pour 2, pour un total de 8 mois» est un compromis raisonnable si le niveau et le changement mensuel des estimations de tendance et des estimations désaisonnalisées sont tous deux considérés comme importants.

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada
Volume 26, numéro 2, décembre 2000

TABLE DES MATIÈRES

Dans ce numéro	131
I.P. FELLEGI Leslie Kish – Une vie de dévouement	133
D.E. HAINES, K.H. POLLOCK et S.G. PANTULA Estimation de la taille et des chiffres de population pour des échantillons tirés de listes incomplètes avec probabilités d'inclusion hétérogènes	135
J.-F. BEAUMONT Une méthode d'estimation en présence de non-réponse non-ignorable	145
B.D. SPENCER Un effet de plan de sondage approximatif pour une pondération inégale en cas de corrélation possible entre les mesures et les probabilités de sélection	153
P.P. BIEMER et J.M. BUSHERY Validité de l'analyse markovienne de structure latente pour l'estimation de l'erreur de classification des données sur la population active	157
K.J. THOMPSON et R.S. SIGMAN L'estimation et l'estimation de la variance par répliques des prix de vente médians des maisons vendues	173
C.H. McLAREN et D.G. STEEL L'effet de divers plans de renouvellement sur la variance d'échantillonnage des estimations désaisonnalisées et des estimations de la tendance	185
Y. YOU et J.N.K. RAO Estimation bayésienne hiérarchique des moyennes pour petites régions à l'aide de modèles à plusieurs niveaux	197
F.C. OKAFOR et H. LEE Échantillonnage à deux phases pour l'estimation par quotient ou par régression avec sous-échantillonnage des non-répondants	207
J. PICKERY et G. LOOSVELDT Modélisation des effets d'intervieweur dans le cas des enquêtes par panel: Une application	213
M. FUCHS La conception des écrans et l'ordre des questions dans un module IAO. Résultats d'une expérience de facilité d'utilisation menée sur le terrain	225
Remerciements	235

LESLIE KISH (1910 - 2000)

Ce numéro est dédié à la mémoire de Leslie Kish. Sa joie de vivre contagieuse, sa préoccupation profonde pour les opprimés et les démunis et ses contributions importantes à la méthodologie d'enquête et aux statistiques ont été, et demeurent, une source d'inspiration pour beaucoup.

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Techniques d'enquête est répertoriée dans The Survey Statistician, Statistical Theory and Methods Abstracts et SRM Database of Social Research Methodology, Erasmus University. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

COMITÉ DE DIRECTION

Président	G. J. Brackstone
Membres	D. A. Binder G. J. C. Hole F. Mayda (Directeur de la Production) C. Patrick
Rédacteur en chef	M. P. Singh, <i>Statistique Canada</i>

Rédacteurs associés

D. R. Bellhouse, <i>University of Western Ontario</i> P. Biemer, <i>Research Triangle Institute</i> D. A. Binder, <i>Statistique Canada</i> C. Clark, <i>U.S. Bureau of the Census</i> J. C. Deville, <i>INSEE</i> J. Eltinge, <i>Texas A&M University</i> W. A. Fuller, <i>Iowa State University</i> J. Gambino, <i>Statistique Canada</i> M. A. Hidiroglou, <i>Statistique Canada</i> D. Holt, <i>Central Statistical Office, U.K.</i> G. Kalton, <i>Westat, Inc.</i> P. Kott, <i>National Agricultural Statistics Service</i> P. Lahiri, <i>University of Nebraska-Lincoln</i> S. Linacre, <i>Australian Bureau of Statistics</i>	G. Nathan, <i>Central Bureau of Statistics, Israel</i> D. Norris, <i>Statistique Canada</i> D. Pfeffermann, <i>Hebrew University</i> J. N. K. Rao, <i>Carleton University</i> L. P. Rivest, <i>Université Laval</i> F. J. Scheuren, <i>The Urban Institute</i> R. Sitter, <i>Simon Fraser University</i> C. J. Skinner, <i>University of Southampton</i> E. Stasny, <i>Ohio State University</i> R. Valliant, <i>Westat, Inc.</i> J. Waksberg, <i>Westat, Inc.</i> K. M. Wolter, <i>National Opinion Research Center</i> A. Zaslavsky, <i>Harvard University</i>
---	--

Rédacteurs adjoints

J.-F. Beaumont, P. Dick, H. Mantel, W. Yung et D. Stukel, *Statistique Canada*

POLITIQUE DE RÉDACTION

Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à faire parvenir le texte rédigé en anglais ou en français au rédacteur en chef, M. M. P. Singh, Division des méthodes d'enquêtes auprès des ménages, Statistique Canada, Tunney's Pasture, Ottawa (Ontario), Canada K1A 0T6. Prière d'envoyer quatre exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

Abonnement

Le prix de Techniques d'enquête (n° 12-001-XPB au catalogue) est de 47 \$ CA par année. Le prix n'inclus pas les taxes de vente canadiennes. Des frais de livraison supplémentaires s'appliquent aux envois à l'extérieur du Canada: États-Unis 12 \$ CA (6 \$ x 2 exemplaires), autres pays, 20 \$ CA (10 \$ x 2 exemplaires). Prière de faire parvenir votre demande d'abonnement à Statistique Canada, Division de la diffusion, Gestion de la circulation, 120, avenue Parkdale, Ottawa (Ontario), Canada K1A 0T6 ou commandez par téléphone au 1 800 700-1033, par télécopieur au 1 800 889-9734 ou par Courriel: order@statcan.ca. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête, l'American Association for Public Opinion Research, la Société Statistique du Canada et l'Association des statisticiennes et statisticiens du Québec.



Ottawa

ISSN 0714-0045

Périodicité: semestrielle

N° 12-001-XPB au catalogue

Février 2001

Tous droits réservés. Il est interdit de reproduire ou de transmettre le contenu de la présente publication, sous quelque forme ou par quelque moyen que ce soit, enregistrativement sur support magnétique, reproduction électronique, mécanique, photographique, ou autre, ou de l'emmagasiner dans un système de recouvrement, sans l'autorisation écrite préalable des Services de concession des droits de licence, Division du marketing, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

© Ministre de l'Industrie, 2001

Publication autorisée par le ministre
responsable de Statistique Canada

DÉCEMBRE 2000 • VOLUME 26 • NUMÉRO 2

UNE REVUE ÉDITÉE PAR STATISTIQUE CANADA

TECHNIQUES D'ENQUÊTE



Ottawa

ISSN 0714-0045

Périodicité: semestrielle

N° 12-001-XPB au catalogue

Février 2001

Tous droits réservés. Il est interdit de reproduire ou de transmettre le contenu de la présente publication, sous quelque forme ou par quelque moyen que ce soit, enregistrativement sur support magnétique, reproduction électronique, mécanique, photographique, ou autre, ou de l'emmagasiner dans un système de recouvrement, sans l'autorisation écrite préalable des Services de concession des droits de licence, Division du marketing, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

© Ministre de l'Industrie, 2001

Publication autorisée par le ministre
responsable de Statistique Canada

DÉCEMBRE 2000 • VOLUME 26 • NUMÉRO 2

UNE REVUE ÉDITÉE PAR STATISTIQUE CANADA

TECHNIQUES D'ENQUÊTE





NUMÉRO 2

VOLUME 26

DÉCEMBRE 2000

UNE REVUE
ÉDITÉE
PAR STATISTIQUE CANADA

N° 12-001-XPB au catalogue

TECHNIQUES D'ENQUÊTE



5577

